Copyrighted Material

FUNDAMENTALS OF SOLID-STATE ELECTRONICS

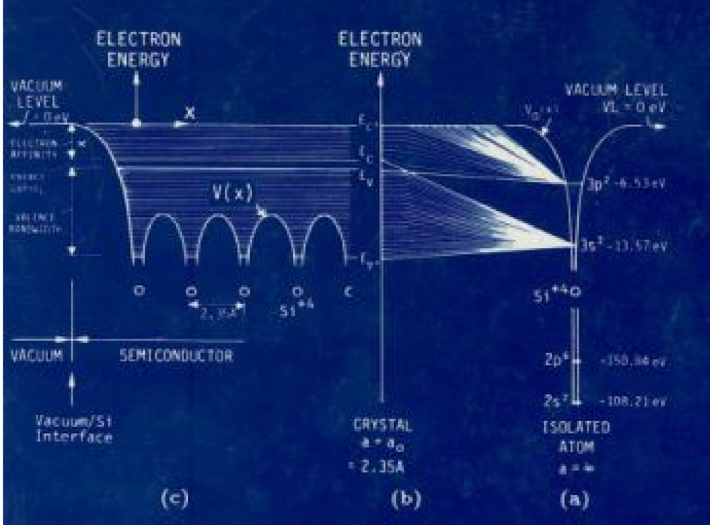

ELECTRON ENERGY
VACUUM LEVEL
X
ELECTRON AFFINITY
ENERGY GAP
VALENCE BANDWIDTH
V(x)
2.35A
Si+4
VACUUM
SEMICONDUCTOR
Vacuum/Si Interface
ELECTRON ENERGY
E_C'
E_C
E_V
E_V'
CRYSTAL
a = a_o
= 2.35A
VACUUM LEVEL
VL = 0 eV
3p² -6.53 eV
3s² -13.57 eV
Si+4
2p⁶ -150.84 eV
2s² -108.21 eV
ISOLATED ATOM
a = ∞
(c)
(b)
(a)

Fig. 172.1

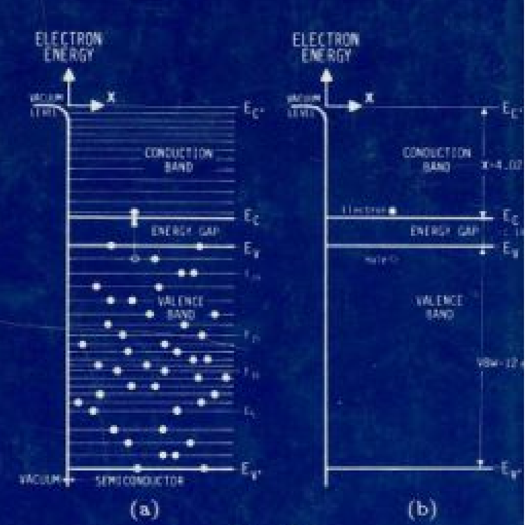

ELECTRON ENERGY
VACUUM LEVEL
X
E_C'
CONDUCTION BAND
E_C
ENERGY GAP
E_V
VALENCE BAND
E_V'
VACUUM
SEMICONDUCTOR
ELECTRON ENERGY
CONDUCTION BAND
electron
ENERGY GAP
hole
VALENCE BAND
(a)
(b)

Fig. 173.2

Chih-Tang Sah

World Scientific

Copyrighted Material

VISIT…

LANZAROTE
Caliente.com

Copyrighted Material

Published by

World Scientific Publishing Co. Pte. Ltd.
P O Box 128, Farrer Road, Singapore 9128
USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661
UK office: 73 Lynton Mead, Totteridge, London N20 8DH

Library of Congress Cataloging-in-Publication Data

Sah, Chih-Tang.
Fundamentals of solid state electronics/Chih-Tang Sah.
p. cm.
ISBN 9810206372. -- ISBN 9810206380 (pbk)
1. Solid state electronics. I. Title.
TK7871.85.S23 1991
621.381--dc20 91-26395
CIP

First printing 1991.
Reprinted 1992.
Reprinted 1993 (pbk).
Reprinted 1994.

Copyright © 1991 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 27 Congress Street, Salem, MA 01970, USA.

Printed in Singapore.

Copyrighted Material

Chapter 1

ELECTRONS, BONDS, BANDS AND HOLES

100 INTRODUCTION 2
110 CLASSIFICATION OF MATERIALS 4
111 Classification Schemes of Solids, 5
* Geometrical Classification (Crystallinity vs Imperfection), 6
* Purity Classification (Pure vs Impure), 6
* Electrical Classification (Electrical Conductivity), 7
* Mechanical Classification (Binding Force), 8
120 CRYSTALLINE AND IMPURE SEMICONDUCTORS ARE NEEDED IN ELECTRON DEVICE APPLICATIONS 10
130 CRYSTAL LATTICES AND PERIODIC STRUCTURES 12
131 Description of Crystal Lattice by Vectors, 13
* Miller Indices, 16
132 Three-Dimensional Crystal Structures, 17
* Diamond, Zinc Blende, Wurzite Structures, 19-22
133 Calculation of the Atomic Density, 22
134 Growing Single Crystals, 23
140 WAVE MOTION OF ELECTRONS IN MATERIALS (Quantum and Wave Mechanics and Schrödinger Equation) 31
141 Dual Character of Material Particles and Electromagnetic Radiation, 32
* The Bohr Model of Hydrogen Atom, 35
* Applications of the Electron Energy Level and Orbit Diagrams, 41
** Emission and Absorption of Light by Hydrogen Atom, 43
** Electron in an Electric Field Around a Proton, 45
* Summary, 46
142 Experimental Bases for Selecting an Equation of the Matter Waves (Derivation of the Schrödinger Equation), 46
143 Properties and Interpretations of the Wavefunction (Classical-Quantum Connections), 53
150 SOLUTIONS OF THE SCHRÖDINGER EQUATION 55
* General Properties, 56
151 Reflection of Electron at a Potential Step, 57
152 Resonance Scattering by a Square Potential Well, 60
153 Tunneling through a Square Potential Barrier, 62
154 Tunneling through a Triangular Potential Barrier, 64
155 Bound States in an Attractive Square Potential Well, 66
156 The Hydrogen Atom, 69
160 ELECTRON CONFIGURATIONS IN MANY-ELECTRON ATOMS 83
161 The Negatively Charged Two-Electron Hydrogen Atom, 83
162 Many-Proton and Many-Electron Atoms, 87
170 ELECTRONIC MODELS OF SEMICONDUCTORS AND SOLIDS 93
171 The Bond Model, 93
172 The Band Model, 97
173 Filling the Energy Band Levels by Electrons, 102
180 ELEMENTARY DERIVATIONS OF THE ENERGY BAND MODELS 108
181 The Nearly Free Electron Model, 109
182 The Tight-Binding Model, 117
183 Energy Band Diagrams of Semiconductors, 125
184 Energy Band of Metals and Conductors, 131
190 COMPLETELY FILLED BAND DOES NOT CARRY A CURRENT (The Concept of Holes) 139
199 BIBLIOGRAPHY AND PROBLEMS 142

100 INTRODUCTION

Passive and **active solid-state** electron **devices** (resistors, capacitors, diodes and transistors) **and integrated circuits** are made up of three types of solids, generally known as **solid-state electronic materials**. They are the **conductor**, **semiconductor**, and **insulator**. Modern solid state electronic devices and integrated circuits usually contain many thin sandwiched layers of conductor, semiconductor and insulator. However, the principal electrical conduction that determines the device and circuit performance occurs in one or more thin layers of semiconductors, hence known as **semiconductor devices** and **semiconductor circuits**. Each semiconductor layer may be composed of a few planes to many million planes of **host** atoms (such as Si, and Ga and As in GaAs), and a few **foreign impurity** atoms (such as B, P, As, Sb in Si).

The analysis, design and manufacture of high performance and reliable semiconductor devices and integrated circuits depend on an in-depth understanding of the fundamental physics and chemistry of these three types of electronic solids.

This and the following two chapters introduce these fundamentals. Chapters 1 and 2 consider **compositionally** or **atomically** and **chemically uniform** electronic solids, at a **spatially and temporally constant temperature**, and not exposed to any applied forces that could disturb the constancy. The constancy of composition and temperature is known as **thermodynamic equilibrium**. It is a global condition consisting of simultaneous **thermal** (constant temperature), **electrical** (zero net electrical current and generation-recombination-trapping of each charged particle species), **chemical** (zero net particle current and reaction from each neutral atomic particle species) and **mechanical** (constant hydrostatic pressure) **equilibriums**. In chapter 3, the fundamentals of electrical and thermal **nonequilibrium** are presented to provide the base for analyzing device and circuit operations since electrical current during device operation causes **static** and **dynamic deviations** from the thermal and electrical equilibrium conditions.

A solid contains many electrons, ions and neutral atoms, about 10^{23} in a volume of one cubic centimeter. There are too many particles and they are too closely packed (small interparticle distances) to be described by classical one-particle and few-particle Newtonian mechanics. In addition, experiments observe and measure only the **average properties** and not the motion of each of the particles. Thus, two extensions of the classical one-particle Newtonian mechanics are made: probabilistic quantum or wave mechanics to deal with the uncertainties from small distances, and statistical mechanics to deal with the large number of particles.

This chapter describes the origin of the quantum mechanical **one-electron model** which gives the allowed one-electron energy states known as energy levels. In solids, these allowed energy levels or states group into bands known as **energy** bands due to the presence of many closely spaced atoms. Derivation and properties of the energy bands are described in this chapter, chapter 1. Chapters 2 and 3 describe the statistical mechanics which show how these one-electron energy states are occupied by the many electrons at thermal equilibrium, and how the occupation is changed by applied forces at nonequilibrium.

The one-electron concept and language of the single electron in an isolated hydrogen atom, and the valence electron concept of molecules, given in freshman chemistry and sophomore physics, can be readily extended to describe the electrical properties of solids due to the many electrons and atoms. Thus, it is essential that the students review the pertinent concepts from freshman chemistry and sophomore physics. For example, the chemical or valence bond model of a molecule from freshman chemistry can be used to provide a vivid **qualitative** understanding of the electrical properties of a many-atom solid. **Quantitative** understanding based on equations and formulae is mandatory in engineering in order to design and manufacture semiconductor devices and circuits. Thus, quantitative concepts and results, learned in **vacuum electronics** or **vacuum physics** from sophomore physics, should be reviewed since they can be extended to solids. These classical concepts include the Newton law of motion (force = rate of change of momentum), the Coulomb law of force between two point charges, and the energy and momentum conservation laws during particle collisions. To apply to the densely packed particles in a solid, the classical concepts in vacuum physics of one or two particles are extended with just two modern quantum and wave concepts, the **Planck** and **de Broglie hypotheses**. **Planck's hypothesis**, also known as the Planck's condition since it was discovered as an experimental condition, states that the energy of an electromagnetic or particle wave, E, is quantized and is given by $E = h\nu$ where ν is the frequency of the wave function that describes the electromagnetic field or the wave motion of a material particle and h is a universal constant known as the Planck constant. **de Broglie's hypothesis** states that the momentum of an electromagnetic or particle wave, p, is inversely proportional to the wavelength, λ, given by $\lambda = h/p$. This quantum and wave mechanical extension of the classical Newtonian mechanics gives the concept of the one-electron energy **band model**.

The 10^{23} electrons and vibrating ions **collide** with each other **continually**, **randomly**, and **rapidly** (10^{13} collisions per second in a solid) due to the electric or Coulomb force between two point charges. In the one-electron energy band model described in chapter 1, these collisions are ignored. In chapters 2 and 3, the effects of these collisions on the electrical properties of the electronic solids are described. The dynamics or motion of an electron in solids under the influence of static and time-dependent electric and magnetic fields differ from that in vacuum because of these collisions, in particular, the **scattering** of the electrons by the vibrating ions in the solid. The principal interaction force between the valence electrons and vibrating ions is the Coulomb force. The strength of the interaction force and its orientation dependence distinguish one semiconductor or solid from another in their electrical, optical, and even mechanical properties. The high rate of scattering is the fundamental cause that puts the electrons in thermal equilibrium with the vibrating host ions. The thermal equilibrium condition is quantitatively defined as the equality of the average kinetic energy of the electrons and ions: KE(electron) $= 3kT_e/2$ = KE(ions of the solid lattice) $= 3kT_L/2$. They are measured by the same temperature, $T_e = T_L$. This thermal equilibrium condition determines the number-distribution of the electrons as a function of their kinetic energy. The number distribution function, known as the **Fermi function**, gives the number of electrons in a given and narrow range of kinetic energy or velocity. It is derived from statistical mechanics. Instead, its derivation based on a very simple application of the kinetic theory is given in chapter 2.

DC and time-dependent electrical current flowing in a transistor or integrated circuit during operation disturbs the electrical and thermal equilibrium. The **particle distribution** in space, velocity, and energy is changed from the Fermi function. In chapter 3, the effects of an externally applied perturbing force are described qualitatively by the valence bond model and quantitatively by the energy band model. The effects are mathematically analyzed and represented by the **Shockley Equations** which are derived from the band model.

The quantitative energy **band model** is supplemented by the **qualitative chemical bond model**. The bond model gives a vivid mental picture which greatly enhances one's understanding of the band model. The band and bond models also introduce a new particle concept, a positively charged electron known as the **hole**. The hole and the electron are known as **quasi-particles** to solid-state or condensed matter physicists. They are **quasi** (or not real) since they are mathematically derived concepts to picture and analyze the properties of the 10^{23} 'physically real' electrons and atoms. **The electrons and holes in the one-electron model are in fact electrons moving in the electrostatic or Coulomb force field of all other electrons and ions present in a solid.** The electrons in solids are quasi-particles (not real) if we call electron in vacuum a real-particle.

The importance of **periodicity** and **randomness** on the characteristics of solid materials and devices are also described in chapters 2 and 3. Several examples are given to illustrate the effects of randomness arising from **physical defects** (physical imperfections due to displaced or dislocated host atoms) and from residual and artificially **incorporated chemical impurities**.

110 CLASSIFICATION OF MATERIALS

Materials are traditionally classified according to their viscosity into **solid**, **liquid**, and **gas phases**. The properties related to viscosity are the diffusivity of the constituent atoms, the atomic density or the number of atoms occupying a unit volume or a volume element, and the hardness or mechanical strength, listed in this or another sequence depending on one's focus to explain and to relate them. Atomic diffusivity rate is low in solid, high in liquid and very high in gas phase.

We would not be able to retain the solid form if atomic diffusivity were high in solids. Similarly, the atomic density is high in solid, medium in liquid and low in gas. The high density leaves small channel openings and gives high interparticle force to retard diffusion. The same can be said about hardness or mechanical strength which is a measure of the cohesive force that binds the atoms together and resists deformation of the material's shape. Experience tells us that solid is hard and maintains a form while liquid and gas do not maintain and would assume the form of the solid container which holds them. But these are all understood in terms of interparticle spacing which determines the magnitude of the interparticle Coulomb force.

A deeper and esoteric classification of the three material phases is the aperiodicity or randomness of the locations of the constituent atoms and molecules. This is measured by an **order parameter** known as the **correlation length** which is a characteristic distance of a material within which distance the atomic positions show a regularity. Thus, a crystalline solid or crystal has long-range order, meaning that the atoms in the crystal are located on a periodic lattice. A non-crystalline solid usually has short-range order, that is, it may be composed of many small crystallites. There may be hundreds or thousands of atoms in each crystallite and the atoms are ordered or arranged on a periodic lattice in each crystallite. The order or crystallinity of a solid can also be destroyed intentionally or unintentionally so that it does not even have short-range order and it may have only local order in each small atomic cluster containing a few atoms (less than five or so). Such a solid is known as amorphous. Liquid has short-range order and the short-range-ordered configurations of its atomic arrangements changes with time within our manual observation period. Glass is a special solid or liquid known as a supercooled liquid which has local or short-range order. The short-range order is frozen into a glass at room temperature or at other observation temperatures when cooled from a higher temperature. So, the local order in a glass will not change with time during our observation period. Molecules and atoms in the gas phase are located completely randomly and their position changes rapidly with time because the separation or inter-atomic and inter-molecular distances in a gas are so much larger than the size of the atoms and molecules. The large interparticle separations in gases diminishes the interparticle forces or the influence of one atom or molecule on the motion of another atom or molecule except when they are atomically close to each other and collide with each other.

111 Classification Schemes of Solids

The solid phase of material is of particular interest since transistors and integrated circuits are made of metals (conductors), semiconductors, and insulators. Several ways of classifying solids according to some specific geometrical, mechanical and electrical properties will be described. Classification or grouping helps to recognize the origins of their unique properties and to invent new solids with specific useful properties.

Geometrical Classification (Crystallinity vs Imperfection)

Solids can be classified according to their **crystallinity** or geometrical perfection measured by the amount of displacement of each atom from its periodic lattice site. A periodic **space lattice** is a three-dimensional array of points formed by the lattice sites with uninterrupted periodicity in each of its three orthogonal directions. A solid is a crystal or is called **crystalline** when all the atoms or molecules in the solid are located on a **periodic** one-, two-, or three-dimensional space lattice. A solid is a polycrystal or is **polycrystalline** if it contains many small and randomly oriented and joined small crystals known as **crystallites**. Crystallites are also known as **grains**, like grains of sand, but grains or crystallites in a solid are joined or held together by some glue. The **interface** or contact surfaces between the crystallites are known as **grain boundaries**. The electrostatic (or **Coulomb**) force due to the charged atoms and electrons in the grain boundary or surface contact layers is the force that glues the grains or crystallites together to form a large polycrystal. It is also the force that holds the atoms together inside the crystallites. A solid is **amorphous** if the atoms and molecules are all randomly located and do not form crystallites.

Thus, a solid is **imperfect** when it is not crystalline or its atoms are displaced from the positions on a periodic array of points. This displacement **imperfection** is known as a **physical defect** or just **defect**.

Purity Classification (Pure vs Impure)

A solid can also be imperfect due to its purity, that is, it contains a variety of **randomly** located **foreign atoms** which are known as **chemical impurities** or just **impurities**. An array of regularly or **periodically** located foreign atoms is known as an **impure crystal** with a **superlattice**. A crystal containing two atomic species each located on its own superlattice is known as a binary crystal. GaAs crystal is a well-known example. If the very-low-concentration phosphorus atoms in an n-type Si were all situated on a periodic lattice, this Si crystal would also be a binary **crystalline** solid with a superlattice. Thus, solids composed of superlattices of different atomic species (Ga and As in GaAs and the hypothetical P in n-type Si) are not defective nor imperfect. However, a superlattice of a low concentration of impurities (such as P in Si) is difficult to fabricate, therefore, the impurities in an impure crystal for electronic applications are usually randomly located. Thus, **impure** crystals are traditionally known as **imperfect crystal**. The effects of the minute amount of impurities on the electrical properties of the solid are mathematically analyzed using a perturbation theory since the position of some of the atoms of the crystal is **perturbed** or **randomly** displaced from its periodic locations.

Thus, a solid can be **physically perfect** or **imperfect**, the latter when its host atoms are displaced from the periodic lattice locations. The solid can also be

chemically pure or **impure**, the latter when it contains foreign impurity atoms. In summary, a solid belong to one of the four categories: (1) perfect and pure, (2) perfect and impure, (3) imperfect and pure, and (4) imperfect and impure. This foursome classification has not been carefully and succinctly delineated in the literature and books. It is the one of two most distinguished and necessary classification schemes to analyze the electrical properties of solids for electronic applications. The other is the electrical conductivity scheme discussed in the next subsection.

In recent and older scientific and engineering literature, the two categories, the randomly distributed physical defects and chemical impurities, are known as **imperfections**. Solids have not been delineated systematically by the four categories probably because their possible natural existence and the technology for their artificial growth were not recognized. For example, perfect and pure solids [category (1)] require the zone refining technique developed in the 1950's and a crystal growth technique that does not produce defects. For a second example, perfect and impure solids [category (2)] require the molecular beam epitaxy (MBE) technique developed in the 1980's. A further probable cause is that their potential applications were not foreseeable. For example, the use of the amorphous hydrogenated Si [impure and imperfect solid in category (4)] to manufacture solar cells for power source applications and as the photoconductor in xerography were demonstrated only since 1985.

The term, imperfection, was first used by John C. Slater in the late 1940's to develop the mathematical perturbation theory of the electrical properties of solids. It was then used in the proceeding title of a 1952 conference attended by the then active and leading solid state physicists, such as Wigner, Slater, Seitz, Shockley, Bardeen, Brattain, Turnbull. See **Imperfections in Nearly Perfect Crystals**, Pocono Pennsylvania, Wiley.

The distinction between physical defects and chemical impurities has not been consistently made by many engineers, physicists, chemists and authors of text and reference books. Instead of using the term **imperfection**, the word 'defect' is used to describe both chemical impurities and physical defects, causing frequent confusion. In this introductory textbook, we will endeavor to make this distinction of terms in order to delineate the physical and chemical differences of the four solid categories and to ingrain into future semiconductor engineers and scientists the fundamental differences.

Electrical Classification (Electrical Conductivity)

In electronic applications, obviously the most important classification parameter is the electrical conductivity or resistivity of the solid. Table III.1 gives the six resistivity or conductivity ranges and the materials in these ranges. The boundaries are not sharply defined due to the temperature dependence of the resistivity. The resistivity or conductivity of a material can vary by many orders of magnitude over the temperature range of operation and cover two or more ranges of Table 111.1. Thus, a solid can be an insulator at a low temperature and a semiconductor at room temperature. The ability to conduct or insulate electrical current, or the figure-of-merit of a solid is also highly dependent on the specific electronic application. Thus, an oxide glass may be an excellent electrical and chemical insulator when used as an outside coating to passivate and protect a silicon MOS VLSI circuit chip against contamination and moisture from ambient in order to function reliably during a ten-year operating life. But the oxide glass makes a very poor electrical insulator for the gate electrode of the MOS transistors inside the DRAM or 80486-CPU chip since oxide glasses would fail (or leak) electrically in a few seconds after a voltage is applied. Both physical defects and chemical impurities play dominating roles in the aging and failure of semiconductor device in these two extremes of application (gate insulator and outside passivation insulator).

Table 111.1
Classification of Solids via Resistivity

MATERIAL TYPE	RESISTIVITY (ohm-cm)	CONDUCTION ELECTRON DENSITY (cm^{-3})	EXAMPLES
A. Superconductor lo-T	0	10^{23}	Sn, Pb at very low temp. ~0.1-4K.
Superconductor hi-T	0	$\propto 10^{23}$	Oxides at higher temp. ~20-100K.
B. Good Conductor	$10^{-6} - 10^{-5}$	$10^{22} - 10^{23}$	Metals: K, Na, Cu, Au.
C. Conductor	$10^{-5} - 10^{-2}$	$10^{17} - 10^{22}$	Semi-metals: As, B, Graphite.
D. Semiconductor	$10^{-2} - 10^{9}$	$10^{6} - 10^{17}$	Ge, Si, GaAs, GaP, InP, $Ga_xAs_yP_z$.
E. Semi-insulator	$10^{10} - 10^{14}$	$10^{1} - 10^{5}$	Amorphous Si.
F. Insulators	$10^{14} - 10^{22}$	$1 - 10$	Diamond, SiO_2, Si_3N_4, Dielectrics.

There are about 10^{23} valence electrons per cm^3 in all solids. Table 111.1 shows that there are fewer conduction electrons. It is the very **inability** of some of the valence electrons to conduct electricity that distinguishes superconductors, good conductors, semiconductors, semi-insulators and insulators. The first three chapters explore the fundamental reasons.

Mechanical Classification (Binding Force)

Kittel, in a most widely used textbook [Charles Kittel, **Introduction to Solid State Physics**, 2nd or 3rd Edition, John Wiley & Sons.], has given a particularly lucid and simple description of this classification scheme at the atomic level based on atomic forces. This is given in the following table. Although this classification is based on the atomic forces that bind the atom together, it puts solids into groups similar to the electrical groups based on the magnitude of electrical conductivity described in the preceding section. This similarity suggests a direct connection between electrical and mechanical properties. The following is a summary. Not only is this classification the principal basis for studying the mechanical properties of solids but it also is intimately connected with the **electrical** and **electronic reliability** of a solid state electron devices, because the chemical ('mechanical') bond that holds the atoms together will act as **electrically charged electron** and **hole trap** if the bond is broken. This will be described further.

A. **Crystal of Inert Gases** (Low temperature solid)
   Van der Wall - London force: dipole-dipole interaction.

B. **Ionic Crystals** (8 to 10 eV bond energy)
   Electrostatic or Madelung force: Coulomb force. NaCl, etc.

C. **Covalent Crystals** (0.5 to 5 eV per bond)
   Electron-pair or homopolar bond. Semiconductors: C,Si,Ge,Sn.

D. **Metal Crystals**
   Delocalized electrons of high concentration, about 1 e/atom.

E. **Hydrogen-bonded Crystals** (0.1 eV bond energy)
   H_2O, Protein molecules, DNA, etc.

The **bond energy** indicated above is a particularly useful fundamental parameter that provides a qualitative gauge on whether (i) the binding force of the atoms in a solid is strong or weak, and (ii) the bond is easy or hard to be broken or ruptured by energetic electrons, holes, ions, and nuclear particles, and by ionizing radiation such as high-energy photons and x-ray. Throughout this book, we use the **electron-volt** as the basic unit instead of joule ($1\text{eV}=1.609\times10^{-19}\text{J}$) used by physicists or kilo-calorie per mole ($1\text{eV}=22.4\text{kcal/mole}$) used by chemists since electrical measurements are made with a voltmeter or oscilloscope in volts. Another rule followed is to choose a unit convention such that the experimental data and fundamental constants are most easily remembered based on basic physics and most easily written down, i.e., in the range of 0.1 to 10.

Bonds in the electronic semiconductors are covalent or slightly **ionic**. Each bond contains two electrons and has been known as an **electron-pair bond**, a concept introduced in 1913 and an electron-dot notation devised in 1916 by the world renowned chemist, Professor Gilbert N. Lewis of the University of California, Berkeley. A bond is **ruptured** or broken when one of its two electrons is removed, for example, by **impact collision** from a fast moving impinging particle such as an energetic electron, a proton, or an atomic ion. The bond can also be broken from exposure to keV electrons or x-ray. These radiations, known as ionizing radiations, will release or eject one of the two electrons in the bond.

After the bond is broken, the host atoms can and will generally be displaced slightly. The displaced host atoms usually cannot return to their original position even if there is a supply of electrons to replace the electron released from the bond because the displacement moves the host atom into a more stable or lower potential energy position. A missing electron in the bond is also a defect since the bond now has only one electron which is known as the **dangling bond**. It has been shown that a dangling bond can drastically affect, especially degrade, the electrical properties and operating life of a silicon MOS transistor and integrated circuit. **Reliability physics** is an engineering science which studies the fundamental atomic mechanisms that cause the electrical properties of a device to degrade or the transistor to 'age' during operation.

120 CRYSTALLINE AND IMPURE SEMICONDUCTORS ARE NEEDED IN ELECTRONIC DEVICE APPLICATIONS

In solid state electronic device and integrated circuit applications, a semiconductor is required which must be crystalline and must contain a carefully controlled concentration (or volume density) of specific impurities. This is the impure and nearly perfect class between categories (3) and (4) in the geometrical-purity classification given in section 111. The slight imperfection comes from the random distribution of the low density of impurity atoms which can be overcome using the MBE technique to give impure crystals with a superlattice, (3). However, MBE is technologically too expensive for production at present (1990). Nevertheless, there are many superior electrical properties (mobility, direct energy-gap for efficient light generation, etc.) possessed by the superlattice which may be desirable in future electronic applications. In this and the next two chapters, we present a detailed description of the physics and mathematical backgrounds which are required to understand the operation of transistors and to design high performance transistors and integrated circuits. A brief summary of the key reasons is given in the following paragraphs.

Semiconductor containing at least two specific impurities is needed for the following two fundamental device reasons:

(1) to provide a **wide range of conductivity** in one semiconductor by controlling its impurity concentration profile (density versus distance or space location) since each group-V impurity atom (P, As, Sb and Bi in Si) gives one negatively charged conduction electron, and each group-III impurity atom (B, Al, Ga, In in Si) gives one positively charged conduction hole, and

(2) to provide **two types of charge carriers** (electrons and holes) to carry the electrical current, or to provide two conductivity types, the **n-type** (conduction by electrons) and **p-type** (conduction by holes).

The wide range of electrical conductivity makes it possible to **control** and **modulate** the magnitude of the conductivity by applying a time-dependent voltage, current, light, temperature, or mechanical force. In contrast, metal has so many conduction electrons that it is difficult to change its **conductivity** by modulating its electron concentration, or its conductance by modulating its conductance channel thickness and length. In the other extreme, insulator has so few electrons that its conductivity and conductance cannot be modulated significantly at all.

Two types of electrical charge **carriers** (electrons and holes) and conductivity types in a semiconductor are required to give high-gain **linear** amplification of electrical signals. Two carrier types are also required to give the highly **nonlinear** current-voltage characteristics necessary to detect, modulate, switch, and waveform shaping or process the electrical signals. The rectification by a p/n junction diode and the signal amplification and switching by a transistor are direct results of the existence of two types of carriers. Even the so-called unipolar field-effect transistors, purportedly to need only one type of charge carrier, could not operate properly and successfully in a circuit application without the two types of carriers and their conductivity. Metals have only one type of charge carriers (the electrons) whose concentration is too high to be modulated by an applied electric field and too high to be changed spatially by impurity doping to give nonlinear current-voltage characteristics. Insulators have essentially no charge carriers.

Crystalline semiconductors or single semiconductor crystals are needed so the defect density is low. Physical defects arise from the random static displacements of the host atoms away from the periodic lattice points. High defect density gives short electron and hole **lifetimes** since defects are **electron** and **hole traps** where the signal-carrying electrons and holes can recombine and disappear. The recombination energy is dissipated as heat or carried away by the vibrations of the host atoms known as **lattice vibrations**. In polycrystalline and amorphous semiconductors, the defect density is very high and the electron and hole lifetimes are very low. Short lifetime means that the electrical signals carried by the electrons and holes are highly attenuated. Polycrystalline semiconductors cannot give high performance solid state electronic devices since the grain boundaries between the crystallites are highly defective or loaded with electron and hole traps for electrons and holes to recombine. Grain boundaries also create blocking potential walls and wells which can prevent the flow of electrons and holes and diminish electrical conduction. Both recombination and blocking are detriments to high signal amplification in a transistor. Amorphous semiconductors are polycrystalline semiconductors with essentially zero grain size and hence have even higher defect densities than polycrystalline semiconductors.

When we discuss the need for two types of impurities in reason (1) previously, we were restricting to the impurities which will only control the electrical conductivity of the semiconductor via controlling the electron and hole concentrations. These are the group V and III impurities in Si stated previously and known as the **dopant impurities**. There is another class of impurity, the group I, II, and VI atoms in the periodic table (such as Au and Ag in Si), whose major effect in Si is to decrease the electron and hole lifetimes. The lifetimes are reduced even when the concentration of the group I, II and VI impurities is very low, at about one part per 10^9 or less. The group I, II and VI atoms in a semiconductor such as Si are known as **recombination impurities** and frequently called the **lifetime killers**. They will change the electron and hole concentrations only very slightly unless their concentration is so high that it is comparable to or exceeds the concentration of the group III and V impurities. In all Si integrated circuits, these recombination impurities are avoided, by purification and careful control of the manufacturing processing steps. Their presence would increase leakage current and standby power. They are not used for lifetime control in high speed integrated circuits since it is not possible to control their densities under the best manufacturing conditions. However, recombination impurities are still used in some discrete (discrete means one transistor or diode on a chip) Si power and switching transistors and diodes to speed up switching in order to decrease power dissipation and increase power conversion efficiency in power control and ac-dc power conversion applications.

In summary, the semiconductor crystal used in electron device or integrated circuit applications must have low defect density and a controlled concentration of selected impurities to control the magnitude and type of the electrical conductivity.

130 CRYSTAL LATTICES AND PERIODIC STRUCTURES

The preceding discussion on physical defects shows that position periodicity of the host atoms have an important effect on the electrical properties of the solid and the performance of the transistors and integrated circuits. Deviation from periodicity due to a displacement of the atoms from their periodic position gives rise to randomness which reduces electron and hole lifetimes and mobilities or conductivities. The **topology** or **geometry** of the periodicity of the host atoms has another important effect on all the properties (electrical, chemical, optical, mechanical) of a solid. For example, the **shape** and size of the periodicity, known as the **unit cell** of atoms, are the key factors that distinguish an insulator and a semiconductor from a conductor or metal; a Si semiconductor from a Ge or GaAs semiconductor. In fact, the shape and the size are determined by a more fundamental factor, the number of protons in the nucleus of the host atom of the crystal. The force that determines the arrangement and the distance between the atoms is the Coulomb force between the positively charged protons in the nucleus and the negatively charged electrons around each atom and at neighboring atoms.

In order to understand and analyze the relationships between the geometrical shape and size of the periodic unit cell and the electrical and other properties of solids, some universal crystallographer's notations and rules will be discussed in the following subsections. These notations and rules are used to catalog and characterize the large number of fundamental crystal lattice structures in terms of the symmetry properties of the geometrical shape and size. Such a systematic approach is necessary since these fundamental structures are the basic building blocks of the infinite varieties of solid materials in the universe.

For example, the three-dimensional crystalline solids found in nature can be grouped into 7 crystal systems, 14 lattices known as Bravais lattice, 32 crystal point groups each with a different point symmetry (such as the rotation symmetry operations about a point), and a total of 230 space groups each with a different space symmetry (or three-dimensional extended symmetry such as mirror reflection in contrast to the point symmetry). To make sense from this large number of groups requires a consensus approach to facilitate analysis and inter-researcher communication. These notations and rules are arrived at from a systematic exploration of the symmetry properties of the shape and size. They can be rigorously derived mathematically using linear algebra and group theory which are given in advanced courses on group theory, solid state physics and solid state chemistry. We shall only give a summary of the several elementary properties which will help us to begin an understanding of the physics underlying the differences in the electrical properties of several semiconductors such as Si, Ge_xSi_{1-x} and GaAs.

131 Description of Crystal Lattice by Vectors

A crystal or crystalline solid is a material whose atoms are situated **periodically** on one or more interpenetrating arrays of points known as a **crystal lattice** which is frequently called the physical lattice, the real lattice or the **direct lattice**. The term, direct lattice, is particularly useful in the mathematical (Fourier) analysis of the position-dependent crystal properties in the **position or physical space**. This Fourier analysis uses the **space harmonics**, $\exp(i\mathbf{k}_n\cdot\mathbf{r})$ where n is the index of the n-th space harmonic. The twin of space harmonics is the familiar **time harmonics** used in the Fourier analysis of time-dependent signals in electrical circuit analysis. The time harmonics is $\exp(i\omega_m t)$ where m is the index of the m-th time harmonic.

The unique physical, chemical and electronic properties of crystals are intimately related and in fact determined by their geometrical features which are **periodic** or repetitive in the physical space. In order to describe these periodic properties quantitatively or mathematically, the following list of concepts and terms have been introduced and universally used by scientists and engineers. They are used to describe both the direct and the reciprocal lattices and we shall give many illustrations.

Unit Cell	(not unique)
Primitive Unit Cell	(not unique)
Basis Vectors a, b, c	(not unique)
Primitive Basis Vectors	(not unique)
Translation Vector of the Lattice	$\mathbf{R}_n = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}$
Primitive Translation of the Lattice	
Miller Indices	

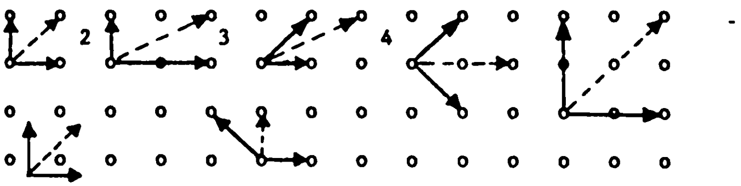


Fig.131.1 Basis vectors and lattice translation vectors of a two-dimensional square lattice.

The variety of **unit cell** is unlimited. The concept and selection of a **unit cell** is illustrated by the seven examples in a two-dimensional square lattice shown in Fig.131.1. The only requirement on constructing or discovering the **unit cell** of a given three-dimensional crystal is that when the unit cell is translated by all the combinations of three integers (positive, negative and zero for n_1, n_2 and n_3) of a corresponding **lattice translation vector**, $\mathbf{R}_n$, of the periodic lattice, the many (infinitely many in the infinitely large crystal) translation operations of the unit cell will precisely fill up all the space occupied by the crystal. The unit cell is obviously **not unique** as the seven alternatives in Fig.131.1 indicate. The unit cell, having the smallest volume in 3-d, smallest area in 2-d, or smallest length in 1-d crystals, is known as the **primitive cell**. The corresponding lattice translation vector is known as the **primitive lattice translation vector**. Obviously from Fig.131.1, the primitive cell is also not unique - four of the seven are primitive, having the same and smallest area of 1 square. Some of the primitive and non-primitive cells offer **higher** geometrical **symmetry** than others and are preferred when a mathematical analysis of the crystal properties is carried out using Fourier series expansion. For example, the primitive unit cell with the atom located at the center of the cell (Exercise: sketch this in Fig.131.1.) was selected by Wigner and Seitz in 1933 in the one of the historical-first mathematical analysis of the electronic properties of a solid, the metallic sodium. [E. Wigner and F. Seitz, "On the constitution of metallic sodium," Physical Review 43, 804 (1933) and 46, 509 (1934).]

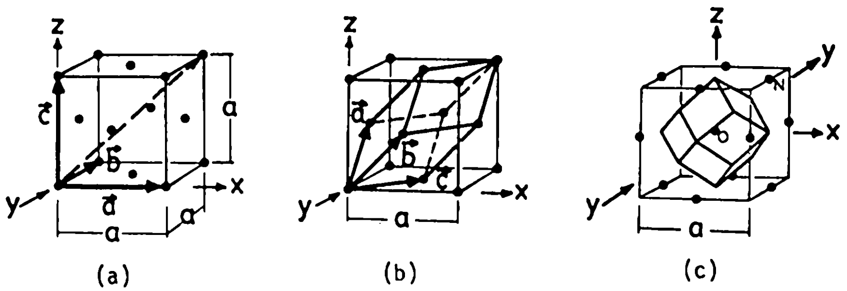
The lattice translation vectors of the crystal can be mathematically represented by one, two or three **basis vectors**, **a**, **b**, and **c**, in the one-, two- and three-dimensional crystals. This is

$$\mathbf{R}_n = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}. \qquad (131.1)$$

The coefficients, n_1, n_2, and n_3 are positive and negative integers including zero. Examples of the the basis vectors (solid arrows) and lattice translation vectors (dash arrows) of the two-dimensional unit cells are shown in Fig.131.1.

Once the crystal structure of a solid is determined, the unit cell and basis vectors can readily be obtained by inspection of the geometry. For example, the simplest primitive basis vectors of the two-dimensional square lattice in Fig.131.1 are $a1_x$ and $a1_y$ and the primitive unit cell is a square. The primitive lattice translation vectors are given by $R_n = n_x a1_x + n_y a1_y$.

The face center cubic (fcc) lattice is now described as a three-dimensional example. Three unit cells are shown in Fig.131.2.



(a) (b) (c)

Fig.131.2 Unit cells and basis vectors of the face-centered cubic lattice. (a) The non-primitive cubic unit cell. (b) The primitive parallelepiped unit cell. (c) The primitive Wigner-Seitz unit cell.

The first is the cubic unit cell shown in Fig.131.2(a). It shows that there is one lattice point (or atom if the solid is monoatomic) at each of the eight corners of the cube and one lattice point (or atom) at center of each of the six square faces. The basis vectors of this cubic unit cell are $a1_x$, $a1_y$ and $a1_z$, and the lattice translation vectors are given by $R_n = n_x a1_x + n_y a1_y + n_z a1_z$. It is not primitive, that is, it is not the smallest volume unit cell as proved below.

The second unit cell of the fcc lattice is the non-cubic but primitive parallelepiped unit cell shown in Fig.131.2(b). The three basis vectors are labeled a, b, and c whose cartesian (x,y,z) components can be readily written down by inspection and is left as an exercise for the students.

The third unit cell of the fcc lattice is the primitive octahedron (eight faces) cell shown in Fig.131.2(c). It has a lattice point (or atom) at the center of the unit cell. A sphere can be inscribed inside this cell. It is known as the Wigner-Seitz cell which was employed by Profs. Eugene Wigner and Frederic Seitz in 1933 when Seitz did his doctoral thesis at Princeton to ease the calculation of the electronic properties of solids. They discovered this unit cell geometry in order to use the spherical coordinate system and the spherical functions to compute the electronic properties of metallic sodium crystal. [See E. Wigner and F. Seitz, "On the constitution of metallic sodium," Physical Review 43, 804 (1933) and 46, 509 (1934).]

Miller Indices

Special sets of notations and symbols have been used by solid state physicists, chemists and material scientists to denote the crystal directions and planes for the three-dimensional crystals. These are known as the Miller Indices whose symbols and notations are listed as follows.

Plane (hkl)	Equivalent planes {hkl}
Direction [hkl]	Equivalent directions <hkl>

The (100), (110) and (111) planes for the cubic lattices are shown in Figs.131.3(a), (b) and (c).

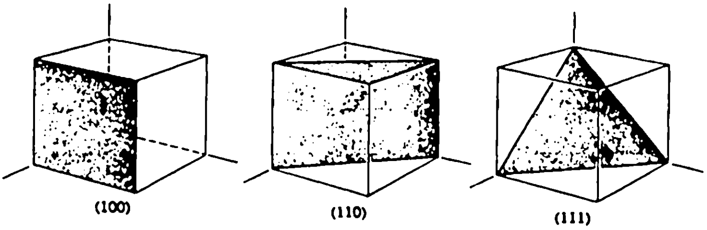


Fig.131.3 Miller indices of the three important planes in a cubic lattice. (a) The (100) plane. (b) The (110) plane. (c) The (111) plane.

The procedure to determine the Miller indices is as follows. The three Miller indices are the smallest set of integers computed from the reciprocal of the intercept of the plane with the three axes. Thus, the (100) plane has the intercepts of $(1,\infty,\infty)$ whose reciprocals are (1,0,0) so that the Miller index of this plane is (100). Similarly, the cube face with $x=0$ in the y-z plane would have intercepts of $(0,\infty,\infty)$ so that its Miller index would be $(\infty,0,0)$. The equivalent (100) planes of the faces of a cubic unit cell whose center is located at the origin of the coordinate are the (-100), (010), (0-10), (001) and the (00-1). These equivalent planes are represented by the symbol {100} and known as the set of equivalent {100} planes.

The crystal directions are denoted by [hkl]. For example, the x-axis is the [100] direction while the negative x-axis is the [-100] direction. The group of equivalent (parallel) [100] directions is denoted by <100>.

The Miller indices for directions can also be used for the two-dimensional lattices. However, the Miller indices for crystal planes are not needed in 2-d and 1-d crystals since there is only one plane in the two-dimensional lattice and no planes in the one-dimensional lattice. Using the square lattice given in Fig.131.1, the x-axis is then the [10] direction while the y-axis is the [01] direction. A 45 degree arrow pointing at 1:30pm would be pointing in the [11] direction.

132 Three-Dimensional Crystal Structures

We shall focus on the three-dimensional solids and crystal lattices since the practical semiconductor, Si, and some of the potentially useful semiconductors (SiC, unstrained Ge_xSi_{1-x}, and GaAs) crystallize in three-dimensional structures. However, there have been a great deal of research activities during the 1980's on electrical conduction in two-dimensional and one-dimensional solids, for examples, the commensurately strained Ge_xSi_{1-x} films, the organic semiconductors, and the high-temperature superconductors. So two- and one-dimensional crystals could find significant applications in the future. The three-dimensional analyses can be applied to the reduced dimension (1-d and 2-d) solids with considerable simplifications.

In the **three-dimensional solids** (3-d) there are

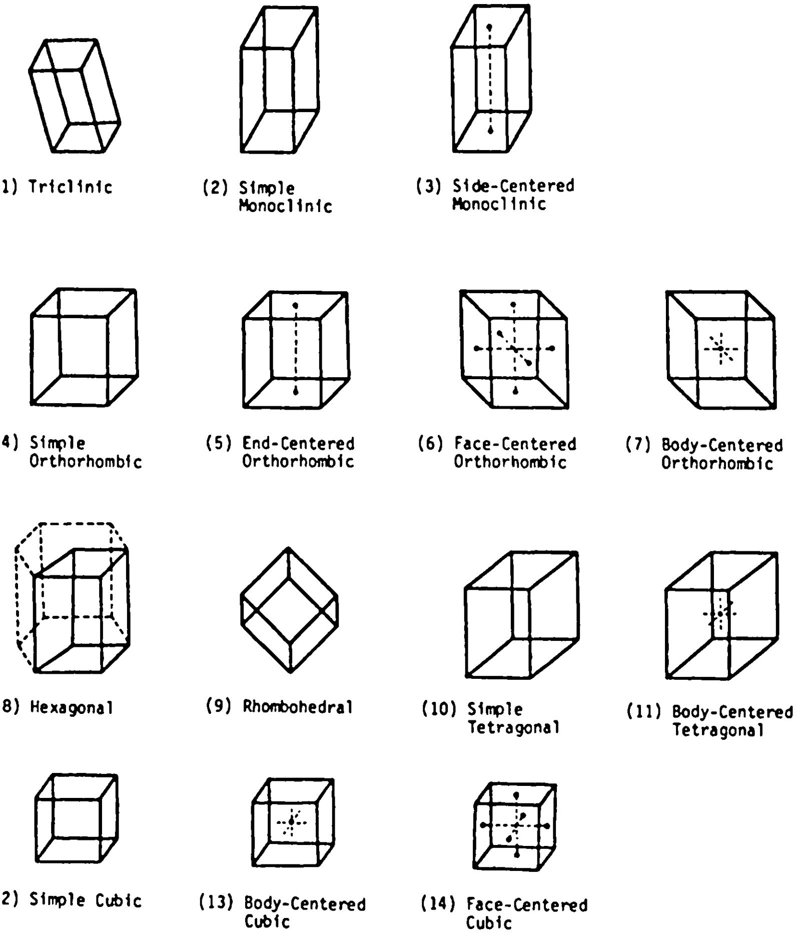
Seven	(7)	3-d crystal systems,
Fourteen	(14)	Bravais or space lattices,
Thirty-two	(32)	crystal point groups, and
Two-hundred-thirty	(230)	space groups.

The seven 3-d crystal systems are (1) triclinic, (2) monoclinic, (3) orthorhombic, (4) hexagonal, (5) rhombohedral, (6) tetragonal and (7) cubic systems. The fourteen 3-d space lattices are known as Bravais lattices and are shown in Fig.132.1. The name of each lattice is given, showing which crystal system it belongs to. For example, there is only one Bravis lattice for the triclinic, trigonal and hexagonal systems; two for the monoclinic and tetragonal systems; three for the cubic system; and four for the orthorhombic system.

The three cubic lattices in the cubic system of the seven 3-d crystal systems are of particular interest since the covalent monoatomic semiconductors, silicon and germanium, the covalent compound semiconductors, SiC and GeSi, and the partially ionic polar compound semiconductors, GaAs, GaP, InP and $Ga_xAl_yAs_z$, all crystallize with cubic symmetry. The hexagonal crystal system is briefly described since some potentially useful compound semiconductors crystallize in the wurzite structure which has hexagonal symmetry. The three basic cubic lattices are

(1)	Simple Cubic	P=Primitive	(SC)
(2)	Body-Centered Cubic	I=Innenzentrierte=body-centered	(BCC)
(3)	Face-Centered Cubic	F=Face-centered	(FCC)

where the abbreviated one-letter and multi-letter acronyms are also given. The German notation came from a 1848 study by the German crystallographer and scientist Adolph Bravais and a 1891-1923 systematic analysis by the German scientist Arnold Schoenflies. A more recent and popular notation is the symbols used in the International Table published in 1935. A comprehensive list was given by the late John C. Slater (former Institute Professor at M.I.T. and Graduate Research Professor at the University of Florida) in a two-volume textbook on solid state theory. The crystal structures of the elemental and compound semiconductors are described in the following paragraphs.



(1) Triclinic
(2) Simple Monoclinic
(3) Side-Centered Monoclinic
(4) Simple Orthorhombic
(5) End-Centered Orthorhombic
(6) Face-Centered Orthorhombic
(7) Body-Centered Orthorhombic
(8) Hexagonal
(9) Rhombohedral
(10) Simple Tetragonal
(11) Body-Centered Tetragonal
(12) Simple Cubic
(13) Body-Centered Cubic
(14) Face-Centered Cubic

Fig.132.1 The fourteen three-dimensional Bravis space lattices.

Diamond Structure (Cubic System)

Elemental Semiconductors: C(diamond), Si, Ge, Sn

The space lattice of diamond is face-centered cubic. Its cubic unit cell is shown in Fig.132.2(a). It is composed of two fcc lattices displaced from each other by one quarter of a body diagonal, (1/4,1/4,1/4)a. This is illustrated in Fig.132.2(b).

The diamond structure has tetrahedral point symmetry (which is one of the five crystal point groups in the cubic crystal system). The tetrahedral symmetry property is also possessed by the other two common crystal structures of semiconductors to be described, the zinc-blende (GaAs) and wurzite (ZnO, ZnS) structures.

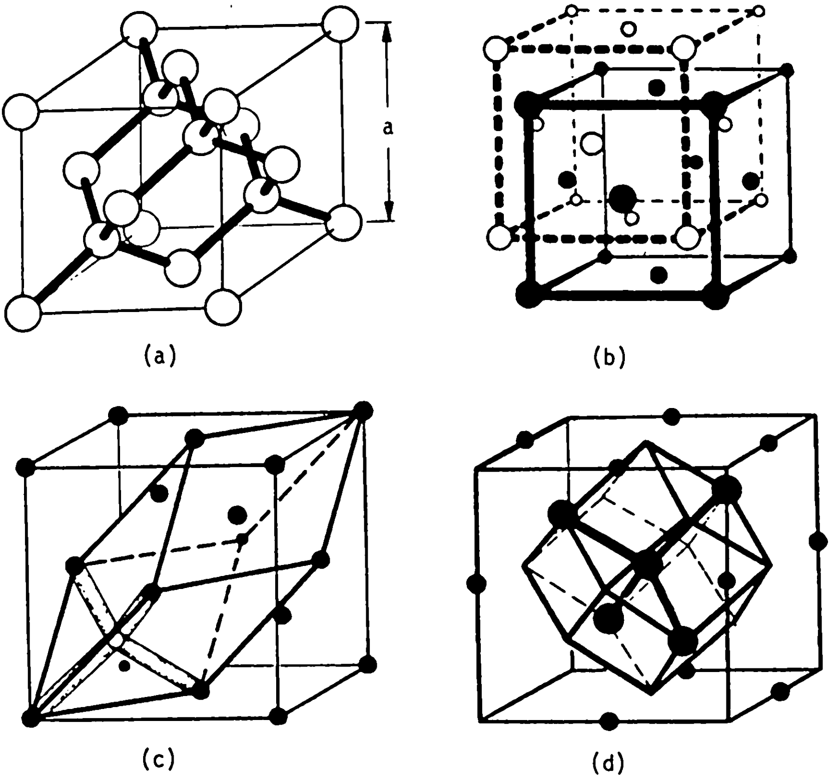


Fig.132.2 The cells of the diamond structure for C, Ge and Si. **(a)** The cubic unit cell. **(b)** Two interpenetrating face-centered cubic lattices. **(c)** The parallelepiped primitive unit cell. **(d)** The Wigner-Seitz primitive unit cell.

The tetrahedral symmetry is illustrated by the small 1/8 size cube inscribed in the cubic unit cell shown in Fig.132.2(a). For the central atom in this 1/8-size cube, there are four atoms located at the four tetrahedral positions relative to the center atom. The four atoms sit at the two opposite corners of the 1/8-size cube.

One particularly important general property of the cubic crystals is that its properties are **isotropic**, i.e., its reaction or response to an applied force is independent of the direction of the force, unless the force is so large that the cubic symmetry is destroyed by the displacement of the steady-state positions of the atoms. In contrast, the **anisotropic** symmetry of the wurzite structure gives very useful electrical and optical properties which are dependent on the direction of the force (mechanical or electrical) and are used to modulate light and to generate electric signals at very precise frequency (quartz oscillator used in electronic wrist watch).

Three examples of unit cells of the diamond crystal structure are shown in Figs.132.2(a), (c) and (d). The cubic unit cell of Fig.132.2(a) contains eight atoms and is not primitive. The parallelepiped unit cell in Fig.132.2(c) and the Wigner-Seitz unit cell in Fig.132.2(d) are primitive and contains two atoms per cell. These three unit cells are identical to those of the face-center cubic lattice shown in Fig.131.2 except that there are twice as many atoms per cell in the diamond structure because there are two interpenetrating fcc lattices in the diamond structure.

Zinc Blende Structure (Cubic System)

Compound Semiconductors
(SiC, SiGe, GaAs, GaP, InP, InAs, InSb, etc.)

Zinc blende crystal structure has the same geometry as the diamond crystal structure except that zinc blende crystals are binary or contains two different kinds of host atoms, such as Ga and As in GaAs. III-V compound semiconductors, such as GaAs, consist of a space array of Ga (or group-III) atoms on one fcc lattice and another space array of As (or group-V) atoms on the other fcc lattice of the diamond structure. This is shown in Fig.132.3(b). Note again the tetrahedral symmetry shown by the 1/8-size cube in the cubic unit cell of Fig.132.3(a). The distinction from diamond is that the four atoms tetrahedrally located with respect to the central atom in zinc blende are different. They may be Ga or group-III atoms while the central atoms would then be As or a group-V atom. Or the four corner atoms may be Si while the central atom is Ge in SiGe and C in SiC.

Wurzite Structure (Hexagonal System Hexagonal Close-Packed Structure)

Compound Semiconductors
(ZnO, GaN, AlN, ZnS, ZnTe,* CdS,* CdTe,* etc.)
(* These also crystallize in the zinc blende structure.)

The adjacent tetrahedrons in zinc blende structure are rotated 60° to give the wurzite structure. This distortion changes the symmetry from cubic to hexagonal and makes the crystal properties **anisotropic**, that is their properties are dependent on the direction of the applied force. The distortion also increases the periodic potential seen by the electrons, making the electron energy gap larger. (Energy gap is discussed later in this chapter.) Figure 132.4(a) shows a larger and nonprimitive hexagonal unit cell while Fig. 132.4(b) shows a smaller and primitive unit cell. The white and black balls in Fig. 132.4(a) represent the two types of host atoms such as Zn and S in ZnS. There has been a renewed research interest recently to grow and characterize several II-VI semiconductors because their larger energy gap offers the potential for solid state diode lasers, optical transistors and photonic devices operating in the deep blue or ultra-violet optical range with wavelength around and below 4000A or 0.4μm. The anisotropy also gives large nonlinear optical properties useful in light modulation applications.

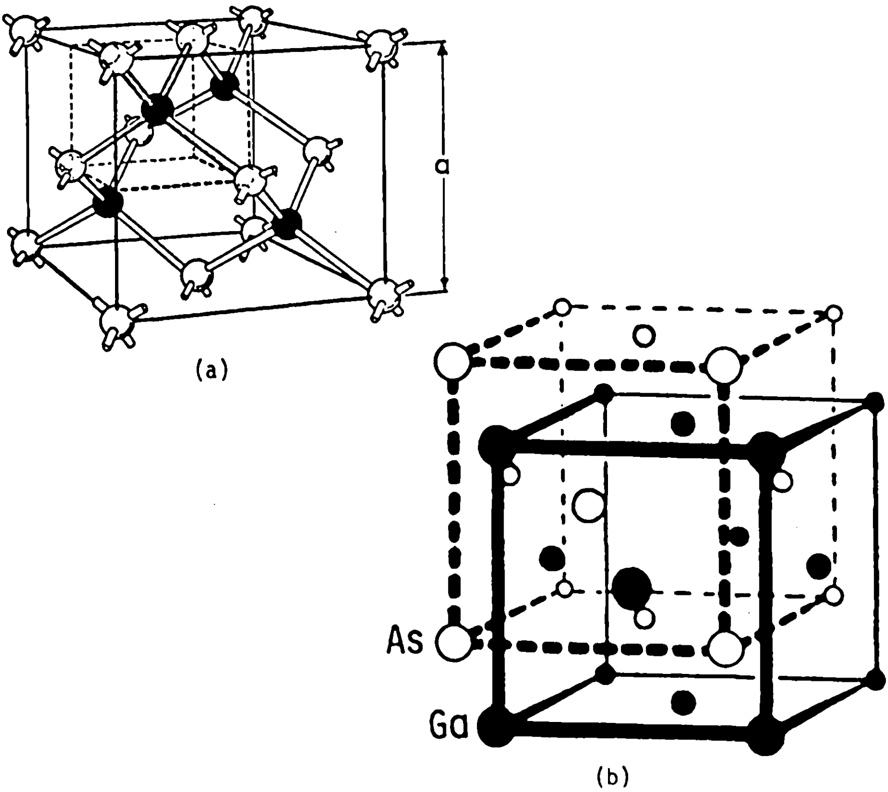


Fig.132.3 The cells of the zinc blende structure. The solid balls are Ga and the circles are As or vice versa. (a) The cubic unit cell. (b) Two interpenetrating face-centered cubic lattices.

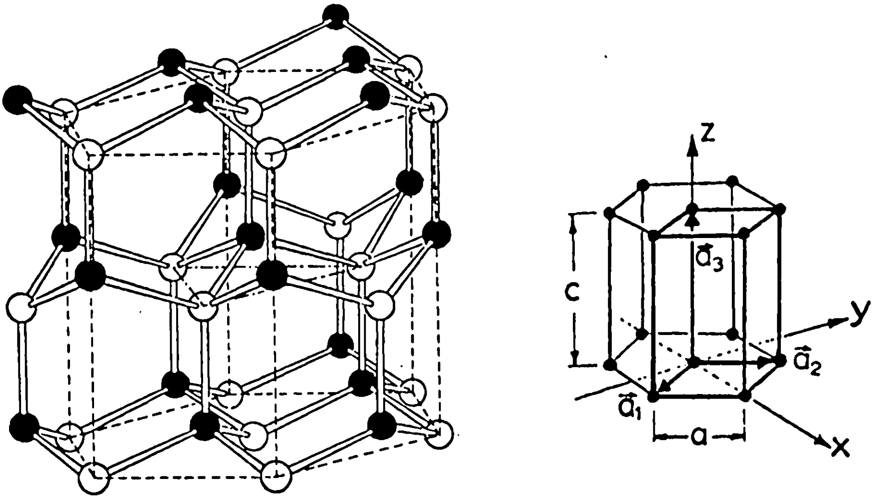


Fig.132.4 The cells of the wurzite structure.

133 Calculation of the Atomic Density

We now work out some examples on how to use the knowledge of crystal structures to compute some useful and important properties such as the number of atoms per unit cell and the atomic density or atoms per unit volume. For example, we will show that there are 8 silicon atoms per cubic unit cell. We shall first work out a simpler example, the two-dimensional square lattice shown in Fig.133.1. The small circle, o, is one atom. The circle with a dot is one enlarged atom which is used to illustrate that the composition of the atomic core also enters into the construction of a unit cell.

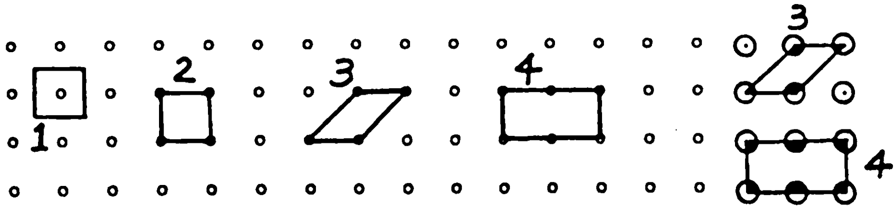


Fig.133.1 A two-dimensional square lattice used to illustrate the calculation of the atomic density of a crystal from a given atomic arrangement or lattice structure and type of host atom.

The following table illustrates the atomic density calculated using four different unit cells. As expected, they all give the same result: the atomic density is

$1/a^2$ where a is the lattice spacing, known as the lattice constant. It is the side of the primitive square unit cell labeled 1.

TABLE 133.1
CALCULATIONS OF THE ATOMIC DENSITY OF A SQUARE LATTICE

Unit Cell Type	Cell Area	Number of Atoms per Cell	Atomic Density
1	a^2	1	$1/a^2$
2	a^2	1 = 4x(1/4) since each corner atom is shared by four adjacent square unit cells.	$1/a^2$
3	a^2	1 by adding shaded areas	$1/a^2$
4	$2a^2$	2 = 4x(1/4) + 2x(1/2)	$1/a^2 = 2/2a^2$

For the diamond lattice such as silicon, refer to the cubic unit cell shown in Fig.132.2(a). The following table illustrates the calculation steps. It gives 8 silicon atoms per cubic unit cell: 8x(1/8) + 6x(1/2) + 4 = 8.

TABLE 133.2
Calculation of the Atomic Density of Silicon Crystal
[Diamond Lattice, Cubic Unit Cell, Fig.132.2(a).]

Unit Cell Type	Cell Volume	Number of Atoms per Cell Calculation
Cubic	a^3	8 corner atoms each shared by 8 cells. 6 face-centered atoms each shared by two cells. 4 atoms inside the cubic unit cell.

134 Growing Single Crystals

In the introduction, the reasons for needing a single crystal or crystalline semiconductor to fabricate transistors and integrated circuits were described. The growth of single crystal is briefly described in this section. Crystals are grown from a liquid or a gas of atoms. There are two essential ingredients to grow a single crystal: an oriented single crystal seed, and a growth condition (temperature gradient, stirring rate and growth rate) to give high mobility (or high velocity) to the atoms so that the atoms have enough time to find and be trapped at a lattice site on the surface of the solid seed before being immobilized by the decreasing solid temperature. The key to optimize the growth condition is to have few physical defects introduced and to have only predetermined and controlled amounts of doping impurity atom incorporated into the crystal to give a predetermined impurity concentration profile (impurity density versus position). The starting silicon can be first purified by chemical means. For example, the elemental and highly pure silicon can be obtained from chemical vapor deposition (CVD) of silicon tetrachloride ($SiCl_4$), trichlorosilane ($SiHCl_3$), or silane (SiH_4) onto a clean substrate or mandril. The resulting Si rods or chunks are polycrystalline. The silicon rod can be further purified by the zone refining method and simultaneously grown into a single crystal. The chunks can also be grown into a single crystal using a horizontal-boat zone refiner or a vertical crystal puller by withdrawing a crystalline seed from a molten Si in a heated crucible (the Czolchralski method). We shall briefly describe these two methods.

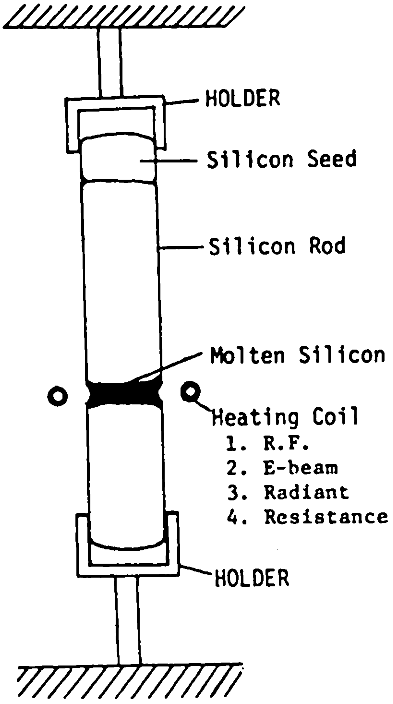


Fig.134.1 A schematic diagram of a vertical zone refining (float zone) apparatus for solid purification and crystal growth known as zone refiner.

The purification of a silicon rod or a boat of silicon chunks by the zone refining method is described first which can also give a single crystal if a crystal seed is placed at one end of the rod. A vertical Si zone refiner is sketched in Fig.134.1, known as the float zone refiner or grower. The whole apparatus is enclosed in a clean enclosure (not shown) and cooled by a water sleeve and/or air (not shown) to keep the interior wall and surface temperature low so that impurities will not outgas from the interior walls and surfaces. The ambient is an ultra pure inert gas, such as helium or argon from a liquid He or Ar tank, to prevent impurity contamination. The purity requirement is so high (as we shall see the reason in chapter 2) that the grade of gases and chemicals used in the manufacturing of the Si transistors and integrated circuits is known as the MOS grade. The MOS grade is purer than the transistor grade, the electronic grade, and a million times purer than the reagent grade use in freshman chemistry laboratory. A crystalline seed is situated at the top (or the bottom) of the polycrystalline Si rod to give a single crystal. Without a seed, the Si rod would still be purified but would still be highly polycrystalline as the original poly Si rod. A thin disk-shaped molten zone of Si, shown in the figure, is created by one of the four heating methods: (1) radio frequency (r.f. at either 450kHz or 1.6MHz), (2) focused electron gun bombardment, (3) radiant focused IR light, or (4) resistance coil. The heating coil moves down, then up, and repeats many times in order to move the molten zone (disk) up and down the rod. One transit of the molten zone over the length of the rod is known as one pass.

This purification and crystal growth process is described as normal freezing or unidirectional solidification. Two key process mechanisms are involved in the first pass of the heating coil traveling down to the bottom from the top of the rod. (1) The polycrystalline rod is crystallized into one single crystal because when the molten zone moves downwards from the Si seed at the top, silicon atoms in the liquid are deposited onto the lattice sites of the cooler (below melting point of Si) Si seed above the molten zone, thereby increasing the length of the single crystal seed and crystallizing the entire poly Si rod into one single crystal rod when the molten zone reaches the bottom holder. (2) The impurities present in the original poly Si rod are swept out into the lower end because the impurities like the liquid better than the solid, namely, they have a higher solubility or are more soluble in the liquid than in the solid.

As the first mathematical example of this textbook, the impurity distribution in a semiconductor rod after one-pass of the molten-zone will be calculated. The analysis uses the coordinate system and symbols shown in Fig.134.2.

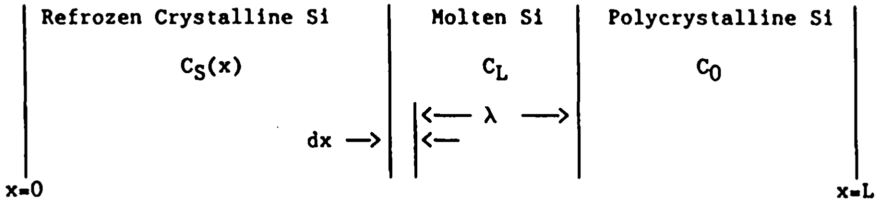


Fig.134.2 Schematic illustrating the calculation of the impurity distribution in a silicon rod after the molten-zone has passed through the rod once.

The following assumptions are made in order to obtain a simplified solution: (1) the thickness of the molten zone, λ, is a constant when it moves from the top end of the Si rod, x=0, to bottom end, x=L; (2) the cross-sectional areas of the rod and the molten zone are equal, constant and given by A; (3) the impurity concentration in the molten zone is spatially constant although it changes with time as the molten zone moves along the rod. Figure 134.1 shows the more realistic geometry of the molten zone.

The total number of impurity atoms in the molten zone of length λ is then $A\lambda C_L$ where C_L is impurity atomic concentration in the molten zone or liquid denoted by the subscript L. C_L is assumed to be spatially constant in the molten zone as stated in (3) above. Then, the change of the impurity atom number in the molten zone comes from two sources: it increases when some polysilicon is melted from the right and it decreases when some molten silicon solidifies onto the refrozen silicon on the left. Thus, material balance or conservation gives us

$$A\lambda dC_L = AC_0dx - AC_Sdx$$

where C_0 is the initial impurity atomic density in the polysilicon rod and is also assumed to be constant. $C_S = C_S(x)$ is the final impurity atomic density in the refrozen or resolidified part of the silicon rod.

C_S and C_L are proportional. The proportionality constant is known as the **impurity segregation coefficient** which measures the amount of segregation of an impurity species between a liquid and a solid solvent in this example. Generally, the impurity segregation coefficient measures the impurity concentration ratio at the interface between two solid solvents, two immiscible liquid solvents, a liquid and gas solvent, a liquid and solid solvent, or a gas and solid solvent. It is defined by

$$k = C_S(x)/C_L(x).$$

If the molten zone moves very slowly, (as compared with the product of the diffusion rate of the impurity atoms in the Si liquid and the attachment rate of the liquid Si atoms to the solidified Si lattice), then k approaches its thermal equilibrium value so it becomes an **equilibrium constant** sometime known as the **distribution coefficient.** Consequently, in the thermal equilibrium limit, it is a fundamental constant of the silicon material and the impurity in question. If the zone moves very fast or if the Si rod on one side of the molten zone rotates rather fast relative to the Si rod on the other side, k is no longer a constant and the mathematics becomes more complex. Rotation is used in crystal growth to produce a more uniform concentration of the impurity in the liquid hence more uniform concentration across the cross-sectional area of the crystal rod. In the float zone method described here, the zone is sufficiently thin so that rotation is not needed to homogenize the impurity in the liquid. In crystal growth by pulling a seed from a crucible of molten Si, to be described as a second example, rotation of the seed is essential to give better uniformity of impurity concentration.

A constant segregation coefficient is assumed in this analysis to simplify mathematics and focus on the basic phenomena. It can then be substituted into the material balance equation above to eliminate one of the two variables, C_S and C_L. This gives

$$(\lambda/k)dC_S/dx = C_0 - C_S.$$

This simple first order differential equation can be integrated from $x=0$ to x, using the initial distribution: $C_S(0)=kC_L(0)=kC_0$. The result is

$$C_S(x) = C_0[1 - (1-k)\exp(-kx/\lambda)]. \qquad \text{(Region I)}$$

This solution is no longer valid in a layer of thickness λ at the end of the rod, labeled region II in Fig.134.3, since there is no solid Si on the far surface of the molten zone. The correct solution in region II is

$$C_S(y) = kC_0[1 - (y/\lambda)]^{k-1}. \qquad \text{(Region II)}$$

The solutions are graphed in Fig.134.3 for $k=0.1$, showing considerable variation of the impurity concentration over the length of the crystal. A nearly constant impurity concentration can be obtained by adding a minute concentration of the desired impurity to the inert ambient gas. The gaseous impurity atoms would rapidly saturate the molten Si zone. However, this gas phase impurity doping technique is controllable only down to about 0.1 atomic percent (one part per thousand) and hence cannot give the low impurity concentrations (one part per million or less) needed to build Si transistors.

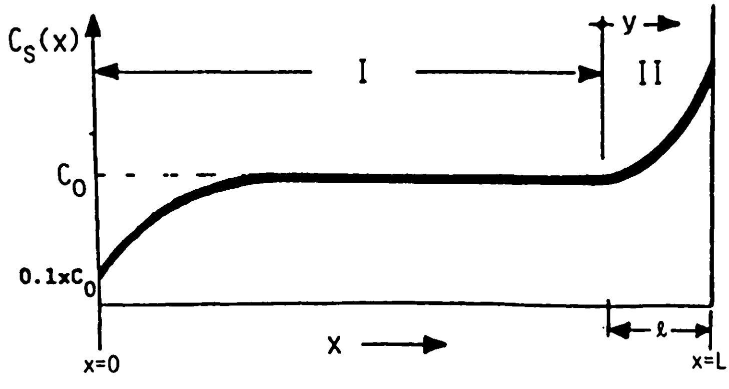


Fig.134.3 The variation of the impurity concentration in a Si rod after one float zone pass.

The second and one of the most common Si crystal growth methods is to pull a single crystalline seed slowly from an impurity-doped molten Si in a carbon crucible lined with a high purity silica (fused quartz). This is known as the

Czolchralski technique and the apparatus is known as the crystal puller. A schematic diagram is shown in Fig.134.4. The parts are labeled and self-explanatory and will be briefly explained.

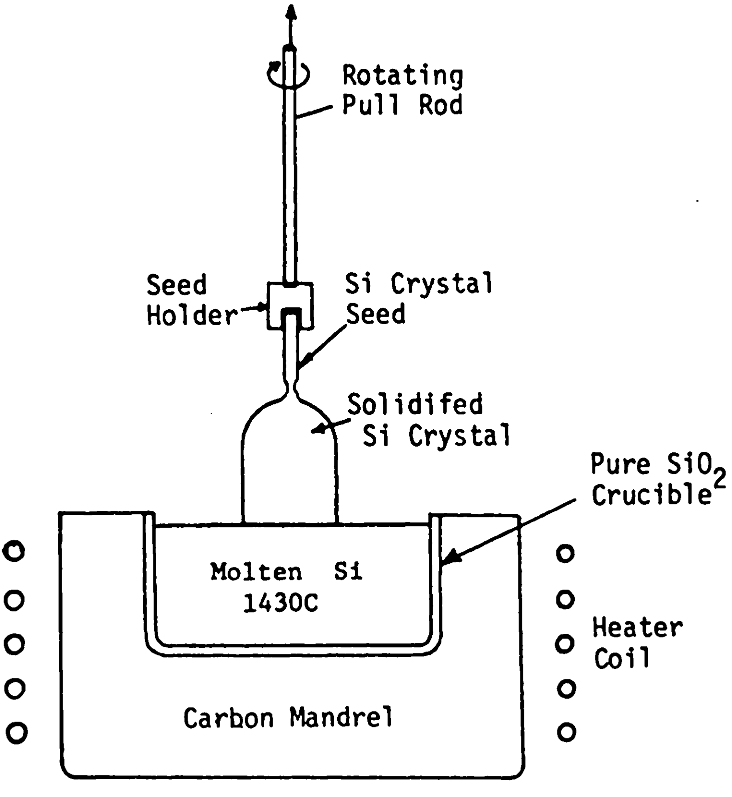

Rotating
Pull Rod
Seed
Holder
Si Crystal
Seed
Solidifed
Si Crystal
Pure SiO_2
Crucible
Molten Si
1430C
Heater
Coil
Carbon Mandrel

Fig.134.4 A Czolchralski crystal puller for growing single crystals.

The entire crystal puller of Fig.134.4 is enclosed by a envelope container made of highly pure fused quartz in order to control the chemical composition of the gaseous ambient and prevent contamination from the laboratory ambient. A high purity inert gas (He or Ar) is used inside of the envelope to surround the hot puller. High vacuum has been used to reduce oxygen contamination from the fused-quartz crucible (described later). In either case, the heated parts of the puller must have very low vapor pressure at high temperatures to prevent impurity evaporation from the hot surfaces which would contaminate the silicon crystal. To meet this purity requirement is quite an engineering feat due to the very high melting point of silicon, 1430°C.

A highly pure silicon crystal seed is rigidly attached to a pull-rod shown in Fig.134.4. The pull rod may be rotational or stationary depending on whether it is necessary to stir the melt in order to homogenize the impurity in the melt. The pulling and rotational motion of the pull rod is controlled by a servomechanism.

The molten silicon is contained in a mandrel made of highly pure carbon and lined with a highly pure fused quartz SiO_2 linear, usually in the shape of a cup. The carbon mandrel is then heated by a radio frequency or resistance power source via the heater coil shown in Fig.134.4. It is evident that the molten silicon is in contact with the SiO_2 crucible. Thus, a substantial amount of oxygen from the crucible is dissolved in the molten silicon via the reaction $SiO_2+Si(liquid) \rightarrow SiO\uparrow+Si+O\rightarrow 2Si+2O$. Some of the reaction product (silicon mono-oxide SiO) will evaporate and deposit onto the cooler interior surfaces of the puller and the quartz envelope. The concentration of oxygen in the solid silicon is essentially equal to the solid solubility limit at the melting point of silicon, about 5×10^{18} oxygen/cm^3. If oxygen is uniformly dispersed in the single crystal silicon, even such a high concentration will not affect the characteristics of transistors and integrated circuits since oxygen is not electrically active, that is, it does not trap electrons and holes. However, high concentrations of oxygen atoms will form SiO_2 particles and clusters around defects or nucleation centers if the crystalline silicon is heated to a high temperature (less than 1420°C) for a prolonged time, such as during transistor fabrication. During this heating, SiO_2 clusters are formed since at temperatures less than the melting point, the oxygen solubility in Si is lower than the maximum solubility at the melting point. Thus, the excess oxygen in the Si must precipitate out onto nucleation sites. Heating to a high temperature during oxidation or impurity diffusion will increase the mobility or diffusivity of the oxygen atoms and speed up their migration towards the nucleation sites.

The SiO_2 clusters are insulating regions which increase the local electrical field if they are situated in the high electric field regions of a Si p/n junction. This high electric field will cause high current to flow which degrades the electrical performance of transistors. The high current in the high electric field region can also charge and generate electron and hole traps. Electrical performance will degrade as traps are charged up or generated, causing transistors to age and integrated circuits to fail. Special heating cycles have been invented and incorporated into the oxidation and diffusion cycles of integrated circuit manufacturing processes. These heating cycles will deplete the oxygen in a thin surface layer (about 10-micrometers and known as denuded layer) of the Si crystal wafer. The heating cycle will also cause oxygen to cluster and SiO_2 particles to form in the interior or bulk region of the Si wafer. The SiO_2 clusters are preferential sinks for metallic impurities which are deterimental to electron-hole lifetimes. Such a process has been widely employed to getter metallic impurities that are present or inadvertently introduced into the silicon wafers during transistor fabrication. Alternatively, oxygen can be avoided using the float-zone crystal grower since the molten zone is not in contact with an infinite source of oxygen (from the SiO_2 crucible liner in the Czolchralski puller). Thus, any oxygen contamination will be swept to the end of the rod or evaporated from the molten zone and deposited onto the cooled wall during the float zone passes.

The basic principle of growth and impurity doping profile control of the Czolchralski crystal pulling method are identical to those of the float zone technique. To start a crystal growth, the tip of the crystal seed is dipped into the molten Si and then withdrawn slowly by a servomechanism controlled stepping motor. The heater is also controlled to keep the molten Si at a constant temperature. The Si and impurity atoms in the molten Si are trapped onto the atomic sites at the Si-liquid/Si-solid interface of the seed.

The concentration of the impurity atoms on the solid Si surface is determined by the impurity segregation coefficient between the Si liquid and Si solid phases, $k=C_S(x,t)/C_L(x,t)$. The solid/liquid interface at at $x=0$, and t is the time from the start of growth. The impurity concentration profile along the length of the crystal, $C_S(x)$ vs x, can be computed from the method just described for the float zone process whose results are shown in Fig.134.3. In the Czolchralski or pulling method, as the seed is withdrawn from the melt (liquid), the volume of the melt decreases. Since k is usually less than 1 because impurities are more soluble in liquid than solid, the impurity concentration in the Si melt increases with time due to the decreasing volume of the Si melt. Thus, the impurity concentration in the Si solid will increase greatly with distance away from the seed end. However, it is immediately obvious that a constant impurity concentration in the Si melt and thereby also in the Si solid can be maintained to give a Si crystal rod with nearly constant impurity concentration over the entire length of the rod if pure Si chunks are continually added to the Si melt in the crucible.

There are several precautions: (1) mechanical disturbances of the molten Si due to adding Si to the melt must be minimized or eliminated, and (2) sufficient time must be allowed for the impurity in the melt to homogenize by diffusion when additional pure Si chunks are added. The chemical cleanliness and mechanical controls required to add Si chunks into a crucible of molten Si at 1420°C are not trivial. These factors have been successfully controlled to give production quantities of Si single crystals (8-inch diameter, several feet long) which are almost physically perfect and have nearly constant dopant impurity concentration. They have been used in megabit DRAMs (1, 4, 16Mb) manufacturing by IBM since 1988.

140 WAVE MOTION OF ELECTRONS IN MATERIALS (Quantum and Wave Mechanics and Schrödinger Equation)

The electrical characteristics of solid state electronic devices and integrated circuits are determined by the motion of many electrons (10^{23}) in the solids. In vacuum, the motion of a **few widely separated** objects (one or two particles or bodies) has been described successfully by **Newton Law** of motion: force = mass x acceleration. Several force rules or **postulates** were established experimentally as **laws**, such as the gravitational force, electric force (**Coulomb Law**) and magnetic force (**Ampere Law**). These classical laws of particle motion in vacuum must be extended in order to describe the motion of the many electrons in solids due to one single fundamental factor that distinguishes solid, liquid, and gas from vacuum. It is the particle density. In solids, there are about 10^{23} electrons and ions packed into a volume of about 1 cm^3 volume. In contrast, there are only about 10^9 electrons in a TV or computer monitor vacuum-tube and 10^{10} electrons in a microwave-oven magnetron vacuum-tube. There are several consequences of this high packing density in solid. (1) The interparticle distance in a solid is very small, about $(10^{23})^{-1/3}$cm = 2×10^{-8}cm. (2) The force acting on the j-th particle comes from all the other 10^{23}-1 particles. (3) Due to the small interparticle distances, the rate of collision between particles is very high, about 10^{13} collisions per second. The high particle density, $10^{23}cm^{-3}$, has given the term, **condensed matter physics** - the fashionable modern name for solid-state physics.

The consequence of small distances stated in (1) is that classical mechanics must be replaced because the instantaneous position of the particle can no longer be defined. In fact, it is no longer meaningful since the particle cannot be seen or located by light without disturbing the particle position when the light is scattered off the particle into a light detector such as the human eye through an optical microscope. Similarly, the velocity or momentum of the particle is no longer **deterministic**. Thus, the electron motion in solids must be analyzed by a **probabilistic theory** instead of the deterministic classical theory of the Newtonian mechanics. To achieve this, the classical mechanics is **extended** using probability, and the new mechanics is known as **wave** or **quantum mechanics**. The differential equation that describes the position probability of a particle is known as the **Schrödinger equation**. In the next several sections of this chapter, we shall review the experimental bases of wave or quantum mechanics and give elementary descriptions of its application to electronic motions in vacuum, in a single atom, and in the many-atom, many-electron solids.

The consequences stated in (2) and (3) make it algebraically unwieldy to derive the equation of motion of the j-th particle using the classical Newton particle-equation or even the modern Schrödinger wave-equation or quantum-equation. However, there is a more practical reason on the inapplicability of the single particle result. This comes from the measurement condition of an experiment we perform. The d.c. value or the waveform of a current or voltage we measure in an observation time interval, such as one second using a d.c. meter or one nanosecond (10^{-9} second) using an oscilloscope, is an average over many collisions of one or many particles. Thus, to compare experiment with theory, to predict experiments by theory, or to design transistors theoretically and then measure the transistor to confirm the theoretical design, we are more interested in the **average electron motion, averaged over many electrons and many collisions**, instead of the motion of each electron at a given instance of time. This experimental condition dictates that the mechanics of motion of the electrons be analyzed by a **statistical** model that facilitates an average over many electrons and collisions. This is known as **statistical mechanics**. **Chapter 2** will describe the results of statistical mechanics analysis at thermodynamic equilibrium, known as **equilibrium statistical mechanics**, which gives the Fermi (or Fermi-Dirac) quantum-distribution of the electron kinetic energy in a solid (condense matter) and the Boltzmann classical-distribution of the electron and particle energy in a gas (dilute matter). **Chapter 3** will describe the results of statistical mechanics analysis at electrical and thermal nonequilibrium when there is an electrical current. This is known as **nonequilibrium statistical mechanics** or **kinetic theory**, which gives the average rate of drift, diffusion, and generation-recombination-trapping of electrons and holes in a solid.

The correctness of wave-quantum (probabilistic) and statistical (averaging) mechanics can only be tested by comparing their theoretical predictions with experiments. In the following section, the inverse is reviewed: the historical experiments that led to the development of the quantum theory and Schrödinger wave equation.

141 Dual Character of Material Particles and Electromagnetic Radiation

Experimental physics before 1900 demonstrated that most of the physical phenomena can be explained by Newton's equation of motion of material particles or bodies and Maxwell's equation of electromagnetic waves and light. These are known as 'classical physics'. For example, the motion of mechanical objects on earth, celestial bodies, and gas molecules could all be predicted by Newton's equation of motion and classical statistical mechanics (the kinetic theory of gases). For another example, the wave nature of light, suggested by Young's diffraction experiments in 1803, could be explained by Maxwell's electromagnetic wave equations which connected the optical phenomena with electrical phenomena. A list of key advances is given in Table 141.1. We shall discuss most of the experiments listed only briefly. For detailed descriptions, the students are to review their freshman chemistry, sophomore physics, and junior modern physics textbooks. Page numbers from three physics textbooks are listed in Table 141.1. We will discuss the basic physics and the consequences of the Bohr atom model in more details. We will give an experimental and physics based derivation of the Schrödinger wave equation and several solutions that are useful in semiconductor device analyses.

Table 141.1
Experiments on the Dual Properties of Material Particles and Light

DESCRIPTION OF EXPERIMENT			Wave or Particle	Page No. of Ref. Books		
				SZY	ER	HR
1803	Young	Diffraction of Light.	W	710-711		994
1912	Laue	Diffraction of Light.	W	729		1057
1900	Planck	Black body radiation, $E=h\nu$.	P	298-301*	8-21	1091
1904	Einstein	Photoelectric effect.	P	756-757	31-38	1096
1923	Compton	Compton effect (X-ray scattered electron in graphite target).	P	769	38-45	1100
1908	Ritz-Rydberg	Combination principle of atomic spectra.	D	757-759	106-109	1115
1907	Einstein	Specific heat of crystalline solid	D	284-285*	421-425	
1912	Debye	Specific heat of crystalline solid	D			
1913	Frank-Hertz	Frank-Hertz experiments, quantized atomic states, electron absorption by gas-vapor.	D		118-121	
1913	Bohr	Bohr atom. Quantized angular angular momentum.	D	759-762	109-118	1105
1922	Stern-Gerlach	Electron spin.	D		296-302	813*
1925	Uhlenbeck Goldsmith	Electron spin.	D			
1924	de Broglie	de Broglie hypothesis, $\lambda=h/p$.	W,P	769	63,77	1117
1925	Pauli	Exclusion Principle.	D	775	334,346	
		Atomic Structure.	D	775-779		
1926	Schrödinger	Schrödinger Equation.	W	(See section 142.)		
1927	Heisenberg	Uncertainty Principle and matrix quantum mechanics.	W,P		72-77,85	1125
1927	Davisson-Germer	Electron diffraction by crystal	W	769	63-70	1118-1120
1928	GP Thomson	Electron diffraction by crystal	W			

W = Wave-like
P = Particle-like
D = Discrete values
SZY = Sears, Zemansky, Young, University Physics, 5th ed. (Freshman-Sophomore General Physics textbook.)
ER = Eisberg, Resnick, Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles, 1974 ed., 10th printing (Sophomore-Junior Modern Physics textbook.)
HR = Halliday, Resnick, Physics, 3rd ed. (1978) (Freshman-Sophomore General Physics textbook.)
* = Description of basic physics is inadequate.

In the next thirty years, many new experiments showed that light behaves like a particle in some experiments while material particles (such as electrons) behave like waves in other experiments. This 'dual character' of light and of material particles cannot be explained by the equations of the classical physics such as the Newton and Maxwell equations. A new equation is needed.

Two fundamental postulates were made to explain the dual character before the new equation was found by Schrödinger in 1926. These are:

(1) **Planck's condition** of absorption and emission of electromagnetic radiation in discrete quanta, known as **photon**: the energy of the photon is given by

$$E = h\nu = \hbar\omega \qquad (141.1)$$

where $\nu = \omega/2\pi$ is the frequency of the electromagnetic radiation (the symbol 'f' is used by engineers instead of 'ν' by physicists for the frequency of a.c. signal or electromagnetic field or wave), and

(2) **de Broglie's hypothesis** on the momentum of a material particle wave: the wave length, λ, is inversely proportional to the classical momentum, p,

$$\lambda = h/p \qquad (141.2)$$

where, h or $\hbar$ ($\hbar = h/2\pi$) is known as the **Planck** constant. It is a universal constant and has the value of

$$\begin{array}{llllll}
h & = 6.6262\times10^{-34} & \text{joule-second} & = 4.1357 & \text{eV-fs} & (141.3A)\\
h/q & = 4.1357\times10^{-15} & \text{V-sec} & = 4.1357 & \text{V-fs} & \\
\hbar & = 1.0546\times10^{-34} & \text{joule-second} & = 0.65822 & \text{eV-fs} & (141.3B)\\
\hbar/q & = 6.5822\times10^{-16} & \text{V-sec} & = 0.65822 & \text{V-fs} &
\end{array}$$

Note that the second set of values have their units in eV (electron-Volt) and fs (femto-second, 10^{-15}s). These are particularly convenient unit to use in solid state theory and semiconductor device physics since the electron-pair bond or covalent chemical bond energies are around 4eV, the power supply voltage used by integrated circuits is 5V, and many fundamental rates of characteristic times are in the fs and ps (10^{-12}s or 1000fs) ranges.

Planck's condition was first suggested by Planck in 1900 to remove the ultraviolet divergence in the theoretical energy density spectrum of black-body radiation predicted by classical theory. It was also used successfully to explain the entire experimental spectrum. Planck's condition was later used by Einstein in 1904 to explain the kinetic energy of electrons emitted from a metal surface during an exposure to light in the photoelectric experiment.

de Broglie's hypothesis of the wavelike property of material particles was suggested twenty years later in 1924 in his doctoral thesis. The wavelike property of particles was confirmed in the electron diffraction experiments in crystals performed by Davisson-Germer (1927) and G.P.Thomson (1928). The particle-like nature of x-ray was demonstrated by the Compton effect (1923). All three experiments were explained using de Broglie's relationship, (141.2).

Detailed descriptions and analyses of the many historical-first experiments on the wave-particle dual properties of material particles and electromagnetic radiation are given in freshman and sophomore physics textbooks. They are listed by page numbers in Table 141.1 and the details are to be reviewed by the students.

Among the experiments listed in Table 141.1, six are considered the first key experiments on the dual character of electromagnetic radiation and material particles. These are: (1) Wavelike property of light: diffraction of light by two slits of Young in 1803 (HR-994) and by crystal of Laue in 1912 (HR-1057); (2) Black body radiation spectral density of Planck in 1900 (HR-1091); (3) Photoelectric effect of Einstein in 1904 (HR-1096); (4) Compton effect in 1923 (HR-1100); (5) Davisson-Germer-Thomson's electron diffraction by crystal in 1927-1928 (HR-1118-1120); and (6) Bohr atom in 1913 (HR-1105). Detailed descriptions of these six experiments are given by Halliday and Resnick in their freshman-sophomore general physics textbook cited in Table 141.1. The page numbers in the 3rd edition are given as (HR-page#). We shall only describe the Bohr model of the hydrogen atom. We will discuss details since they are the foundation of the energy band and valence bond theories used in transistor and semiconductor device physics.

The Bohr Model of Hydrogen Atom

Bohr formulated the hydrogen model in 1913, known as the **Bohr Atom**, with four postulates **before** the Schrödinger equation was formulated in 1926. Two of these are fundamental postulates which are items 1 and 4 listed below. The algebraic steps and the physical basis of each step are given in the following four numbered paragraphs.

In Bohr's model, the proton or hydrogen nucleus is situated at the origin of the coordinate system selected, r=0 or r=0. The position of the electron is denoted by the vector r whose magnitude is r. The free or rest mass of electron is m. The electron velocity is v with magnitude v. The magnitude of the electron charge is q. The permittivity of free space or vacuum is ε_0 ($=8.854\times10^{-14}$F/cm).

1. **Orbital angular momentum is quantized.**

$$mvr = n\hbar \quad \text{where } n=1,2,3,4,\dots \qquad (141.4)$$

There is a very simple geometrical illustration of this assumption. Furthermore, it is a derived form of the de Broglie hypothesis, which was formulated in 1924, eleven years after Bohr's hypothesis. The geometrical interpretation is that the electron wave is a standing wave around the nucleus. Thus, the circumference of the orbit must be an integer number of wave length,

$$\text{circumference} = 2\pi r = n\lambda.$$

But, de Broglie hypothesized that $\lambda = h/p = h/mv$ where $p = mv$ is the linear momentum. Thus, $2\pi r = n\lambda = nh/mv$. This gives $mvr = nh/2\pi = n\hbar$ which is exactly Bohr's quantum condition for the angular momentum, (141.4). Thus, de Broglie's hypothesis, proposed 11 years after the Bohr atom, was just a generalization of Bohr's quantization condition used in the Bohr hydrogen atom model.

2. **Classical electrostatic force law and Newton's law of motion are valid between the positively charged nucleus (proton) and the negatively charged orbiting electron, both treated as point charges. (Classical mechanics.)**

$$\begin{aligned} &\text{Coulomb force} &&= q^2/4\pi\varepsilon_0 r^2 \\ = \ &\text{Centripetal force} &&= mv^2/r \end{aligned} \qquad (141.5)$$

This Coulomb law applies to two point charges. The radius of the proton and electron, about 1fermi = 10^{-13}cm, is much smaller than the electron orbit radius, 0.53×10^{-8}cm. The proton and electron radii were unknown during Bohr's time. They were measured later. Thus, the point charge assumption on the Coulomb force between the proton and electron in the hydrogen atom is consistent with experiments.

Postulate 1, combined with the Coulomb law given above, limits the electron orbit radii to a discrete set of values known as the **stationary** or **non-radiating orbits**. Combining (141.4) and (141.5), these discrete radii are:

$$r_n = 4\pi\varepsilon_0\hbar^2 n^2/mq^2. \qquad (141.6)$$

The radius of the smallest orbit is given by $n = 1$ and has the value of

$$r_1 = 0.53 \times 10^{-8} \text{ cm} = 0.53\text{A}.$$

This is known as the Bohr radius and frequently denoted by the symbol, r_B and r_0. We shall use the symbol r_1 for the Bohr radius.

3. **The total energy of the electron is the sum of kinetic and potential energies: $E = T + V$. It is computed as follows using Newton's law of force and acceleration.**

$$\begin{aligned} T &= \text{Kinetic energy of the orbiting electron} &&= mv^2/2 \\ V &= \text{Potential energy of the orbiting electron} &&= V(r) \\ dV(r) &= -\mathbf{F}(r)\cdot d\mathbf{r} \quad \text{(derived from Newton's law)} && \end{aligned}$$

where

$$\mathbf{F}(r) = -(q^2/4\pi\varepsilon_0 r^2)\mathbf{i}_r$$

and

$$d\mathbf{r} = \mathbf{i}_r dr.$$

Integrating the potential energy function, V(r), over the entire space to a distance r from the origin where the hydrogen nucleus or proton is located, then

$$\int_\infty^r dV(r) = V(r) - V(\infty) = -\int_\infty^r \mathbf{F}(r)\cdot d\mathbf{r}$$

$$= -\int_\infty^r \mathbf{F}(r)\cdot d\mathbf{r} = +\int_\infty^r (q^2/4\pi\varepsilon_0)(dr/r^2)$$

$$= -q^2/(4\pi\varepsilon_0 r) = -q^2/[4\pi\varepsilon_0\sqrt{(x^2+y^2+z^2)}]. \qquad (141.7)$$

where we have expressed the radial distance in the spherical coordinate, r, by its Cartesian components, x,y,z, using $r=\sqrt{(x^2+y^2+z^2)}$.

$V(\infty)$ is the reference level of the potential V(r) or total energy E of the electron. Since the Coulomb potential energy varies as 1/r, the most convenient choice of the reference potential energy or total energy is $V(\infty)=0$. This is also known as the **vacuum level** of the electron energy level diagram and it is used extensively in the energy band theory of semiconductors and solids. A graphical illustration of V(r) in the y=z=0 plane, V(x,y=0,z=0), is shown in the potential energy diagram given by Fig.141.1(a).

Using the quantized or discrete Bohr orbit radii given by (141.6), the electron potential energy at the distance r_n from the proton (at r=0) is then

$$V(r_n) - V(\infty) = V(r_n) = -q^4m/[(4\pi\varepsilon_0)^2\hbar^2n^2]. \qquad (141.7A)$$

Using the Bohr radii, (141.6), and the quantum postulate of the angular momentum, (141.4), the kinetic energy from Newton's law, $T=mv^2/2$, then becomes quantized or discrete and is given by

$$T(r_n) = mv^2/2 = +(1/2)\cdot q^4m/[(4\pi\varepsilon_0)^2\hbar^2n^2] \qquad (141.7B)$$

whose magnitude is exactly 1/2 of the potential energy, (141.7A). Adding the kinetic and the potential energy, the total energy is then given by

$$\begin{aligned} E &= T + V \\ &= -q^4m/[2\hbar^2(4\pi\varepsilon_0)^2n^2] = E_n = -13.6/n^2 \ \text{(eV)} \end{aligned} \qquad (141.8)$$

This result shows that the total energy of the orbiting electron is **quantized**, that is, it can have only **discrete values**. This quantization comes from Bohr's postulate of quantized or discrete angular momentum in (1) discussed above. The discrete energies are given by -13.6eV for n=1, -3.40eV for n=2, -1.51eV for n=3,.... The negative sign indicates that the total energy is below the top of the Coulomb potential energy well of the electron. Thus, the electron is bound to the well. It cannot escape unless given a kick by some external agent, a force or a particle such as another high-velocity electron or a photon in order to gain sufficient kinetic energy so that the total final energy of the electron exceeds the lid of the well, $E>0$.

These results are illustrated graphically in Fig.141.1(a) to (c) to give the students a feel of the orbit size, and an exercise on the use of the electron energy diagram to be described in the following subsection. Figure (a) shows a cross-sectional view of the potential energy of the electron in the hydrogen atom, plotted as a function of the radial distance between the electron and the proton. The proton with a charge of $+q$ is located at the origin, $r=0$. The potential energy of the electron at the top of the Coulomb well has the reference value of $V(\infty)=0$. This plot is identical to the cross-sectional plot of $V(x,0,0)=-q^2/[4\pi\varepsilon_0|x|]$ which is less confusing since r is measured from the origin radially outward and hence can have only positive values while x extends towards both the left and right and can be either positive and negative values as the figure indicates.

Figure 141.1(b) is the one-dimensional **electron energy level diagram.** The thick horizontal lines are the allowed energies, E_n, at which an electron can be captured by a proton. The line length is the diameter of the orbit. Each of the allowed total energy contains two components, the kinetic and the potential energies. The kinetic energy component of E_1 is +13.6eV. The potential energy component is -27.2eV. Adding them up, the total energy E_1 is -13.6eV.

From the formulae just derived, it is evident that the electron orbit radius, r_n, is half the well radius at the (well potential) energy E_n, i.e. $r_n = r_{2n}/2$ where r_{2n} is the radius of the well potential energy at the energy E_n given by $V(r=r_{2n})=E_n$. This is illustrated by the length of the heavy dark line in the 1-d energy diagram of Fig.141.1(b). It is also illustrated by the radius of the circular orbits in the 2-d x-y plane of Fig.141.1(c) for n=1, 2 and 3. The picture of an electron orbit smaller than the well radius at each allowed energy, E_n, is consistent with the physics that the electron is bound inside the well by the positively charged proton. The orbit picture in the 2-d x-y plane is known as the **electron orbit diagram.**

The two diagrams just described for isolated hydrogen atom (the energy level and electron orbit diagrams) are the fundamental building blocks of the **electron energy band** and **electron bond diagrams** in a solid. They are crucial in the qualitative understanding and quantitative design of the characteristics of transistors and integrated circuits. We shall give additional examples following the next subsection on the electron-photon interaction in excited hydrogen atom.

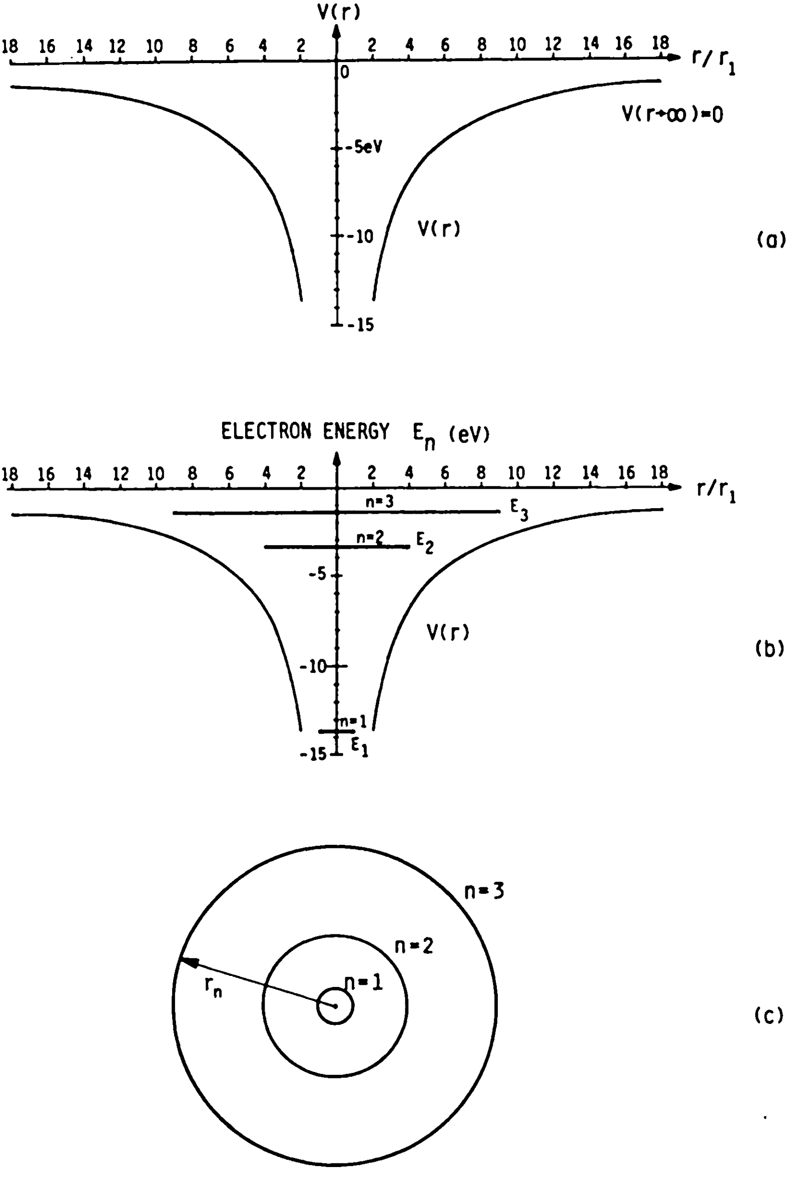

V(r)
18 16 14 12 10 8 6 4 2
2 4 6 8 10 12 14 16 18
r/r1
0
V(r→∞)=0
-5eV
-10
V(r)
-15
(a)
ELECTRON ENERGY En (eV)
18 16 14 12 10 8 6 4 2
2 4 6 8 10 12 14 16 18
r/r1
n=3
E3
n=2
E2
-5
V(r)
-10
n=1
-15
E1
(b)
n=3
n=2
n=1
rn
(c)

Fig.141.1 (a) The cross-sectional view of the potential energy of an electron in the field of a proton, $V(r)$. The proton is located at the origin, $r=0$. The potential $V(r=\infty)$ is selected as the reference potential energy, i.e. $V(r=\infty)=0$. This is known as the vacuum level, VL. (b) The one-dimensional energy diagram of the discrete energies of the electron bound to the potential well of the proton in a hydrogen atom. (c) The two-dimensional electron orbit diagram of the electron bound to the proton.

4. Use Planck's condition to give the energy, frequency or wavelength of light radiated by an excited hydrogen atom.

When the electron bound to the proton in the hydrogen atom is in a high energy orbit, $n>1$, both the hydrogen atom and its bound electron are said to be in an **excited state**. The electron in the excited state $n>1$ will drop down to the lowest energy orbit, $n=1$, known as the **ground state** and a photon is emitted. By the law of energy conservation, the energy of the photon is the difference of initial, E_n, and final, E_1, energies of the electron, $E_{photon}=E_n-E_1$. The frequency and wavelength of the light quanta or photon is obtained from Planck's condition (141.1), giving

$$h\nu_{n1} = E_n - E_1 = 13.6[1 - (1/n^2)]. \text{ (eV)} \qquad (141.9A)$$

The energies or wavelengths of the light or photons emitted by a volume of excited hydrogen gas are given by an infinite series rather than one discrete value. The reason is that there are many excited hydrogen atoms in a volume of hydrogen gas and each at a different excited energy state owing to collisions which transfer and exchange the kinetic energy. These hydrogen atoms come from breaking up the hydrogen molecules in the volume of hydrogen gas, $H_2 \rightarrow 2H$, by impact collision, heating, or some applied forces such as a very intense electric field. This series of color of light or photon energy is known as the **Lyman series**. The first Lyman line of light, known as Lyman alpha line, has a photon energy of $(n=2)$ $h\nu_{21}=E_2-E_1=10.2\text{eV}$.The continuum of the Lyman series starts at $h\nu=E_\infty-E_1=13.6\text{eV}$ and extend to higher energies since some electrons may have received high kinetic energy $(E>0)$ during the collisions. The photons in the Lyman series are in the invisible far ultraviolet range which cannot be detected by the photo-sensitive molecules in the human eye. They were observed as a series of sharp lines when the light from an excited hydrogen gas is viewed on a phosphor screen through a grating spectrometer in a vacuum chamber. Vacuum chamber is employed so that the light is not absorbed by the gas molecules in the ambient. For this reason, they are known as the **vacuum ultraviolet light**.

The next series of light emitted by a volume of excited hydrogen molecules comes from electrons dropping down to the $n=2$ orbit from the higher energy orbits $(n>2)$. This was observed first historically in 1885 by J. J. Balmer because it is in the visible spectrum. It is known as the **Balmer series**. A spectrometer with a thin slit at its entrance and a glass prism is used to analyze the frequency components of the light. The prism refracts the slit-defined line of hydrogen-emitted light into a series of physically separated lines each with a single color or photon wavelength because the glass prism refracts the light of different wavelengths by a different angle owing to the wavelength dependence of the index of refraction of glass. The lines are recorded by a glass plate coated with a photosensitive emulsion. From Planck's condition, the photon energy is given by

$$h\nu_{n2} = E_n - E_2$$

$$= -13.6[(1/n^2) - (1/2^2)] = 3.40[1 - (2/n)^2] \text{ eV}. \qquad (141.9B)$$

The light wavelength can be computed from the frequency using the velocity of light in vacuum, $\nu\lambda = c = 3.0\times10^8$m/s which gives $\lambda(\mu m) = 1.24/h\nu$ where $h\nu$ is in eV. The color of light for the first four and visible lines of the Balmer series are

$n=3$,	$h\nu_{32}=1.89$eV,	$\lambda=0.656\mu m=656nm=6560A$	RED
$n=4$,	$h\nu_{42}=2.55$eV,	$\lambda=0.486\mu m=486nm=4860A$	BLUE
$n=5$,	$h\nu_{52}=2.85$eV,	$\lambda=0.434\mu m=434nm=4342A$	Bluish-PURPLE
$n=6$,	$h\nu_{32}=3.02$eV,	$\lambda=0.410\mu m=410nm=4103A$	Deep PURPLE
$n=\infty$,	$h\nu_{\infty 2}=3.40$eV,	$\lambda=0.364\mu m=364nm=3636A$	Dark PURPLE-invisible

The next series, with final state at $n=3$, is known as the Paschen series. It falls in the infrared range and is again invisible like the Lyman series in the vacuum ultraviolet range. However, its short wavelength limit is at 1.51 eV. This coincides with the infrared light emitted by GaAs diode laser and LED (light emitter diode) and hence has some technological interest.

In the following subsection, the absorption and emission of light by a hydrogen atom from **interaction** of its electron with photon is illustrated by the electron energy level and orbital diagrams shown in Figs.141.1(b) and (c). These two diagrams provide the mental pictures that greatly enhance the understanding of the atomic phenomena and the fundamental physics.

Applications of Electron Energy Level and Orbit Diagrams

The **electron energy level diagram**, Fig.141.1(b), is a one-dimensional picture designed to illustrate the **position variation** of the **total electron energy**, E, in the **physical** or **direct space**. It shows the spatial variation of the **kinetic** energy, T(r), and the **potential** energy, V(r), even if the total energy is a constant. Details are labeled in Fig.141.2(a) for the electron trapped at the bound state $n=2$ by a proton which is located at $r=0$ or $x=y=z=0$. This energy-distance diagram shows the electron energies (E, T, and V) as a function of x in the $y=0$ and $z=0$ plane. The potential energy of the electron are the two heavy curves labeled V(x,0,0) and given by $V(x,0,0) = -q^2/[4\pi\varepsilon_0\sqrt{(x^2+y^2+z^2)}] = -q_2/[4\pi\varepsilon_0|x|]$. As an example, the kinetic energy, $T(x_2,0,0)$, and potential energy, $V(x_2,0,0)$, of the electron located at $(x_2,0,0)$ are labeled in this figure. It is evident that even though the total energy is a constant, i.e., $E = T(x_2,0,0) + V(x_2,0,0) = \text{constant}$, T and V individually varies with the electron position, x_2. (See Problem P141.2.)

The above illustrates the energy diagram of a bound electron. A similar application can be made for a not-bound electron. The electron is not bound to the proton if its total energy is positive, or $E - V(\infty) > 0$. In this case, the kinetic and potential energy of an electron vary with its position. Figure 141.2(b) gives the kinetic and potential energy of the unbound electron located at x_1 near the proton at $r=0$. It is evident that $T(x_1)$ and $V(x_1)$ of the electron vary with its position x_1. The variation comes from the variation of V(x,y,z) due to the positively charged proton.

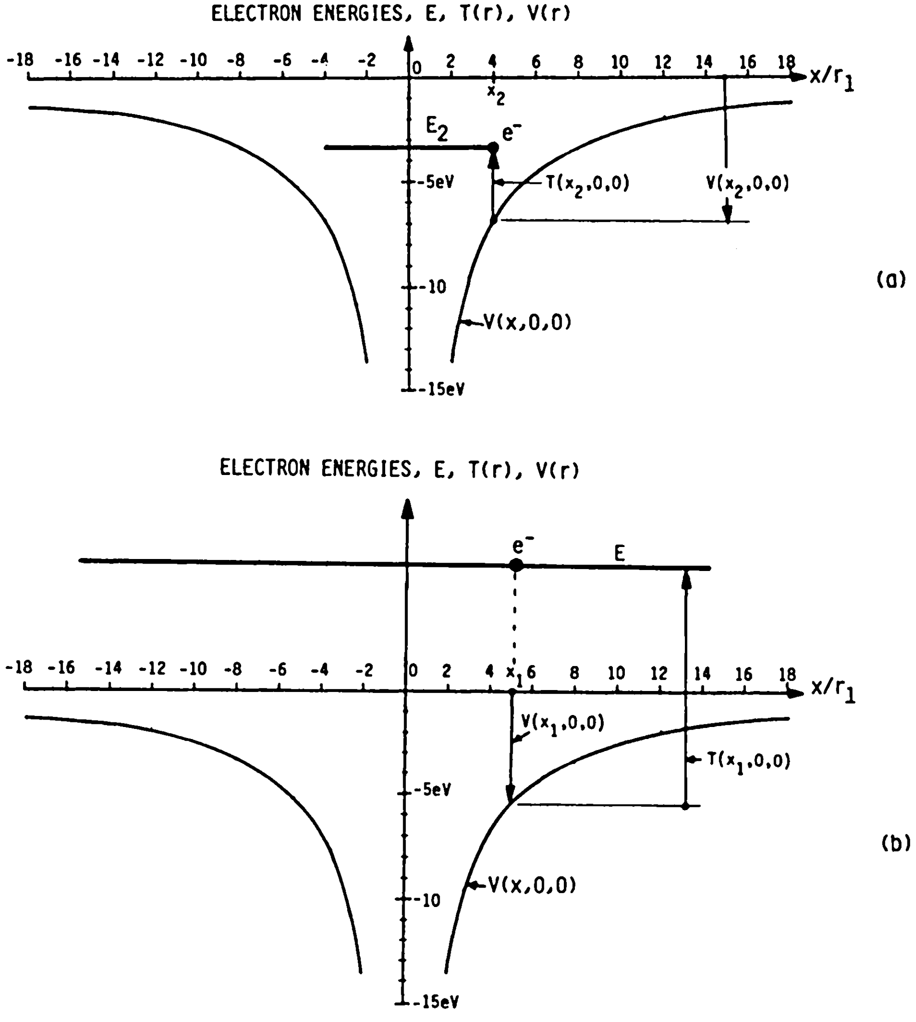


Fig.141.2 The total, kinetic and potential energies of an electron around a proton. (a) The electron is bound to the $n=2$ state of the hydrogen atom. (b) The electron energy is positive $E>0$ and hence not bound to the proton.

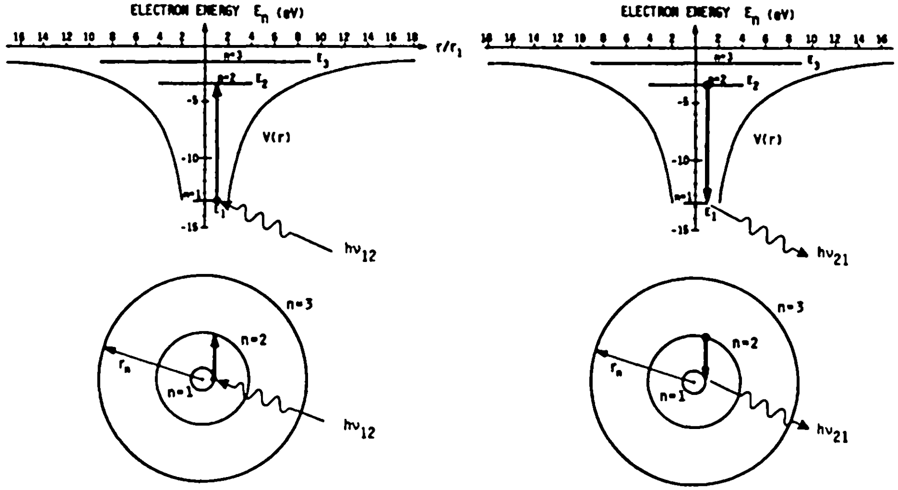
The 2-d **electron orbit diagram**, Fig.141.1(c), is also very useful. It illustrates the collision mechanisms that scatter the electron and change the electron energy. The total electron energy, E, is spatially constant if the collision does not cause the electron to gain or lose energy. This is known as an **elastic collision**. The light absorption and emission collisions just described for the electron bound in the hydrogen atom are **inelastic collisions** since the electron gains energy during a photon absorption transition or loses energy during a photon emission transition. The photon supplies or carries away the electron energy change. Collision of the electron with the vibrating host atoms in molecules and solids is another inelastic collision. These collisions control the electron mobility and lifetime, hence the transistor performance. Thus, the electron orbit diagram, Fig.141.1(c), can help to visualize the collision process, hence the collision rates and the magnitude of the mobility and lifetime or the transistor performance.

Two examples are given to illustrate the use of the 1-d electron energy diagram and the 2-d electron orbit diagram. These are: **electron-photon interaction in a hydrogen atom, and electron motion near a proton in an applied electric field.**

Emission and Absorption of Light by Hydrogen Atom

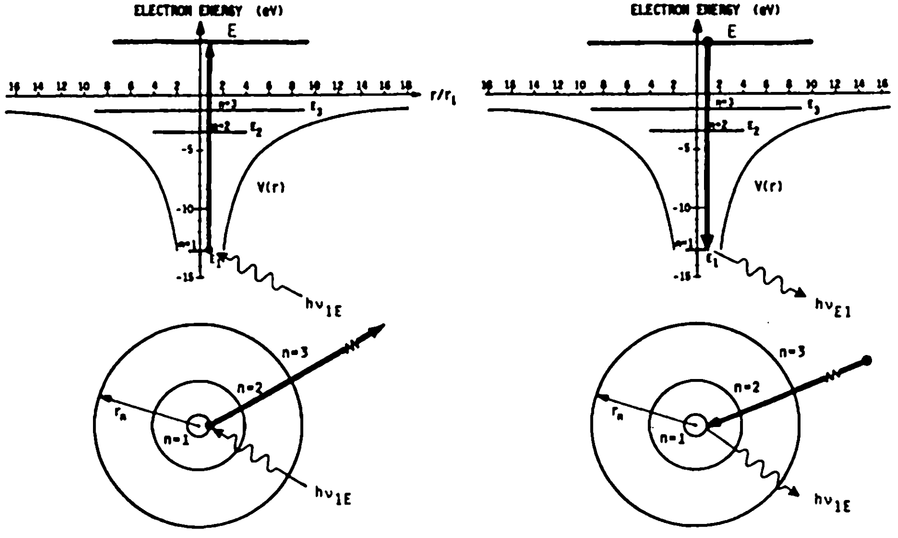
Consider the emission and absorption of light by the hydrogen atom which we have just described without using a figure. We shall now use the 1-d electron-energy and 2-d electron orbit diagrams shown in Figs.141.3(a) to 141.4(b) to explain the electron-photon interactions. Consider first the **absorption of light**. When the hydrogen atom is shielded from all external forces, including the light under consideration, the electron bound to the hydrogen atom will seek the lowest allowed energy level so that the total energy of the hydrogen atom is at the minimum. This is $n=1$ and $E_1=-13.6\text{eV}$ and the electron is represented by a dot at E_1 in Fig.141.3(a). The hydrogen atom is said to be at the **ground state** and the electron is at the ground state or ground energy level.

If the hydrogen atom is now given some energy by exposing it to: an intense electric field, some energetic or high velocity particles (electrons, protons, ions), or some high energy photons (greater than about 10.2eV), two events of the same origin can occur to the bound electron at E_1. (i) The bound electron can be excited to a higher-energy and larger-radius orbit with $n>1$ such as the orbits with $n=2$, 3, or larger and the hydrogen atom is said to be in an **excited state**. (ii) The electron can be stripped off or **released** (emitted or ejected) from the proton and the hydrogen is now **ionized**. The excitation process, (i), by light is shown in Fig.141.3(a) for a photon energy of $h\nu_{12}=E_2-E_1=10.2\text{eV}$. The photo-ionization process, (ii), is shown in Fig.141.4(a) with $h\nu_{21}>13.6\text{ev}$. The one-dimensional **energy level diagram** is shown in the upper part and the two-dimensional **orbit diagram** is shown in the lower part of the two figures. The excitation and ionization processes are depicted by the **electron transition arrow**. The **tail** of the arrow is the initial state of the electron and the head of the arrow is the final state. Evidently, the final state of the ionization transition in Fig.141.4(a) is at $r=\infty$ since $r_n(n=\infty)=\infty$ and the length of the arrow is infinite.


ELECTRON ENERGY E_n (eV)
r/r_1
n=3
E_3
n=2
E_2
-5
-10
-15
n=1
E_1
V(r)
$h\nu_{12}$
$h\nu_{21}$
n=3
n=2
n=1
r_n

(a) (b)

Fig.141.3 Application of the electron energy and electron orbit diagrams to the optical transitions in an excited hydrogen atom. (a) Photon absorption. (b) Photon emission.


ELECTRON ENERGY (eV)
E
r/r_1
n=3
E_3
n=2
E_2
-5
-10
-15
n=1
E_1
V(r)
$h\nu_{1E}$
$h\nu_{E1}$
n=3
n=2
n=1
r_n

(a) (b)

Fig.141.4 Application of the electron energy and electron orbit diagrams to the optical transitions in an ionized hydrogen atom. (a) Photon absorption. (b) Photon emission.

Emission of light by an excited or ionized hydrogen is the inverse of the absorption of light just described and it can also be illustrated by the 1-d electron energy and 2-d electron orbit diagrams. Figure 141.3(b) shows the emission of light when a hydrogen atom in the $n=2$ excited state decays to the ground state $n=1$, giving off a photon of energy $h\nu_{21}=E_2-E_1=10.2$eV. Figure 141.4(b) shows the emission of light when an ionized hydrogen atom in the state $n=\infty$ is neutralized by capturing an electron with energy $E>0$. The photon emitted has an energy of $h\nu=E-E_1 = (E+13.6)$eV. Photons with a continuum of photon energy beginning at 13.6eV will be emitted if there are many electrons having different energies from $E=0$ to higher energies, and if there are many ionized hydrogen atoms or protons which can capture these electrons.

Electron-photon interaction just described is one of the many mechanisms that can scatter an electron. Others are the scattering of an electron by another material particle or the **electron-(material particle) interaction**, such as the scattering of the electron by: a randomly moving molecule in a gas, a randomly vibrating atom in a solid, and a fast moving (energetic) electron, ion, or nucleus.

Electron in an Electric Field Around a Proton

In transistor design, one needs to know the electric field which is given by the gradient of the electron potential energy divided by the electron charge, q, $E=\nabla V/q$.

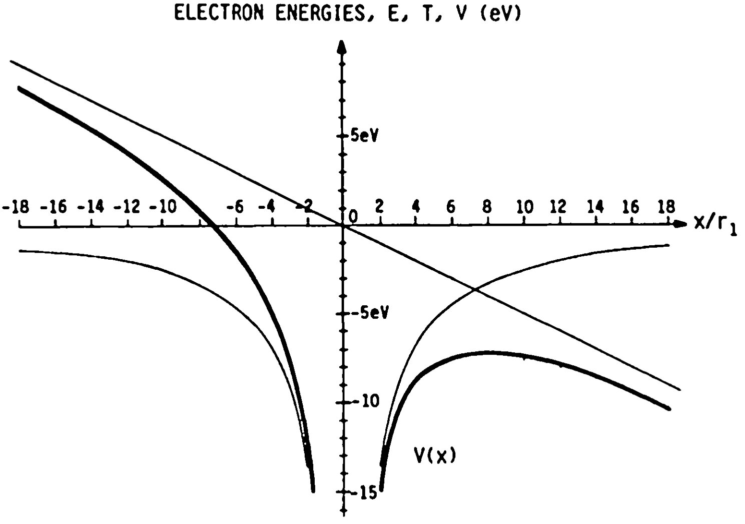


Fig.141.5 The electron energy diagram around a proton in an applied constant electric field.

This connection between the electric field and the energy diagram is an extremely important result in designing transistors and in understanding how transistors work. It will be used repeatedly in the applications of the electron energy diagram to the analysis of semiconductor devices and to the understanding of device physics.

Figure 141.5 shows the electron energy diagram in the presence of a proton, which gives $V_{proton}(r) = -q^2/(4\pi\varepsilon_0 r)$ and a spatially constant applied electric field pointed in the x-direction, $E(r)=E_X i_x$. The constant electric field gives an additional potential energy to the electron which is $V_{applied}(x) = -qE_X i_x \cdot (-x) = +qE_X x$. Thus, the total potential energy is

$$
\begin{aligned}
V(x) &= V_{proton}(r) + V_{applied}(x) \\
&= -q_2/(4\pi\varepsilon_0 r) + qE_X x \qquad (141.10)
\end{aligned}
$$

Summary

In summary, the two extremely useful and important consequences of using the electron energy level and orbit diagrams are: (1) to understand by graphs the fundamental physics of the bound and free electron states, and the collision or transition processes of electrons, and (2) to illustrate the spatial variation of the electric potential and the electric field experienced by the electron since the electron potential energy is just the negative electron charge, -q, times electric potential. These diagrams for an isolated atom are the bases for the energy band and valence bond diagrams of the many-atom solids which are used to explain the properties of semiconductors and solids in sections 17n, 18n, and in chapters 2 and 3, and to explain and analyze the transistors in later chapters.

142 Experimental Bases for Selecting an Equation of the Matter Waves
(Derivation of the Schrödinger Equation)

The fundamental postulates of Planck and de Broglie can only predict some of the simplest results of experiments on material particles. A general theory is needed which not only will encompass the two fundamental postulates but also explain every facet of the detailed results of an entire experiment. It must also be able to predict new phenomena and results of new experiments. The success of the Maxwell partial differential equation of space-time to explain electromagnetic and light wave phenomena had suggested that a similar partial differential equation may be found to describe the wave phenomena of material particles. This differential equation of space-time for material particles is known as the **Schrödinger equation** discovered by him in 1926.

Textbook authors have followed several routes to 'derive' or arrive at the Schrödinger equation. It can be simply written down as a postulate and its correctness is then tested by comparing its theoretical predictions with experiments. This is a common approach since it is concise and short, requiring only three algebraic lines as indicated below. It is known as the **operator approach** or differential operator approach.

(1) $E = (T + V)$

(2) $E\psi = (p^2/2m + V)\psi$

(3) Let E and p be differential operators defined by

$$E \rightarrow +i\hbar\partial/\partial t$$

and

$$p \rightarrow -i\hbar\nabla = -i\hbar[\mathbf{i}_x(\partial/\partial x) + \mathbf{i}_y(\partial/\partial y) + \mathbf{i}_z(\partial/\partial z)]$$

then the Schrödinger equation is obtained and given by

$$i\hbar\partial\psi/\partial t = -(\hbar^2/2m)\nabla^2\psi + V\psi.$$

However, we still need to find the connection between the dependent variable or the theoretical solution, $\psi(r,t)$, and the experimentally observed quantities such as energy, velocity, wavelength, and other material properties.

The foregoing formal approach is neat but it hides much of the interesting physics and it lacks a basis to connect the theoretical solutions to measured quantities from an experiment. The only physics used was: $E=T+V$ or the total energy is equal to the sum of the kinetic and potential energy (of a material particle).

We shall now give a derivation based on a more physical (i.e. physics-based), satisfying, and historical route. It brings out more physics which should help to understand the logic, to remember the physical bases, and to apply the solutions to new and practical problems in transistor engineering and device physics. It follows the time-honored approach of using experimental results to guide the development or 'derivation' of a theory.

Consider first the simple one-dimensional traveling mono-frequency (or single-frequency, **monochromatic**) harmonic waves as a possible solution for the matter waves without knowing as yet what the differential equation is. The possible forms are $\sin(kx-\omega t)$, $\cos(kx-\omega t)$, $\exp[i(kx-\omega t)]$ and $\exp[-i(kx-\omega t)]$ or some linear combination of these where the **wave number** k and **frequency** ω are given or precisely known. Note, $\exp[\pm i(kx-\omega t)] = \cos(kx-\omega t) \pm i\sin(kx-\omega t)$, thus, the complex exponential waves are already linear combinations of the sine and cosine waves. These represent waves traveling from the left to the right or in the positive x-axis direction (why and why not also pick those traveling in the negative x-direction). If these wavefunctions are used to represent the position probability of a particle, then the particle would have a completely undetermined position, since the probability of finding the particle, $\psi^*\psi$, is a constant independent of position. Thus, the particle is equally likely to be found at all locations. But, the particle would have precisely known energy, E, (or kinetic energy since the potential energy is zero) and precisely known momentum p, since ω and k are given. From Planck's hypothesis, the particle energy is

$$E = \hbar\omega \qquad \omega = 2\pi\nu \qquad \nu = \text{frequency}$$

From de Broglie's hypothesis, $\lambda = h/p$, and using $\hbar = h/2\pi$ and $\lambda = 2\pi/k$ from EM or any wave theory, the particle momentum is then

$$p = \hbar k \qquad k = 2\pi/\lambda \qquad \lambda = \text{wavelength}$$

where k is the wave number. Thus, E and p are determined when ω and λ are given in a known wave such as $\sin(kx-\omega t)$. Therefore, these simple sine and cosine waves are too restrictive to represent the results of a general atomic or solid state experiment although they can account for and describe some experiments using classical laws and simple postulates, such as, (i) The photoelectric effect via Planck's postulate of quantization of light or electromagnetic energy, $E = h\nu$; (ii) The Compton effect (wavelength shift of scattered x-ray by electrons in a carbon target) using both the Planck and de Broglie postulates; (iii) Black body radiation spectral density using Planck's postulate; and (iv) Line spectra of light (discrete light frequency or wavelength) emitted by an excited hydrogen atom using Bohr's postulate that the orbital angular momentum of the electron in the hydrogen atom is quantized.

A **partial differential equation**, whose independent variables are space-time, could remove these limitations as suggested by the highly successful previous theoretical experiences on describing many nature and artificial phenomena by partial differential equations. However, we must first decide on the physical meaning of the dependent variable (or the wavefunction) which is the solution of the partial differential equation. We previously pointed out that a probabilistic theory is necessary not only for the single-particle problem such as an electron in a hydrogen atom but even more so in solids where the interparticle distances are very small because there are very many electrons and atoms.

The **first possible choice** of the dependent variable or the solution of the differential equation is the position probability density function, $P(x,y,z,t)$, defined by

$$P(x,y,z,t)\Delta x\Delta y\Delta z\Delta t.$$

It is the total probability of finding an electron (or particle) in the volume element $\Delta x\Delta y\Delta z$ located at (x,y,z) and in the time interval Δt at a time t relative to some reference time $t=0$. It normalizes to unity in order to represent certainty when it is integrated over all space, i.e., a particle is found in all space or exist at all times. This choice is too restrictive since the probability density, $P(x,y,z,t)$, must be a positive real scaler. In addition, it is too complicated mathematically to construct or find a positive real scaler solution to describe the experimentally observed interference phenomenon of two material-particle waves, such as the experiment on electron diffraction by a crystal performed by Davisson and Germer in 1927 and G.P. Thomson in 1928. In these experiments, the incident and diffracted electron waves interfere with each other to give an array of bright spots of high electron density on the emulsion film of the detector.

To seek a less restrictive or more general independent variable or wavefunction, we backtrack, by noting that $P(x,y,z,t)$ is a magnitude or positive real scaler. Thus, the **second possible choice** is to let $P(x,y,z,t)$ be the magnitude of an arbitrary complex function, $P(x,y,z,t) \equiv |f(x,y,z,t)|$ and f must satisfy a partial differential equation. This choice of restricting to a magnitude leads to tedious algebra since we need to compute the magnitude of a complex variable or function using the steps of $|f(x,y,z,t)| = \sqrt{|f^2(x,y,z,t)|} = \sqrt{f^*(x,y,z,t)f(x,y,z,t)}$ where f^* is the complex conjugate of f.

It is immediately evident from the above algebraic steps that a better and **third possible choice** for the dependent variable or the solution of the differential equation is the function itself, $f(x,y,z,t)$ instead of its magnitude (choice 2) or the square of its magnitude (choice 1). This preferred and **consensus third choice** is denoted by $\psi(x,y,z,t)$ and is known as **the wavefunction**. Then, the probability density is given

$$P(x,y,z,t) \equiv \psi^*(x,y,z,t)\psi(x,y,z,t) \equiv \psi^*(r,t)\psi(r,t)$$

which is positive and real, while the wavefunction, $\psi(x,y,z,t)$ is an arbitrary function and a solution of the differential equation for a specific experiment. This choice shows that $\psi(x,y,z,t)$ can be arbitrary, positive, negative, and complex with real and imaginary parts. It can have a phase factor in the form of $\exp(i\theta)$ or $\exp(i\omega t-ikx-iky-ikz)$ demanded by the electron diffraction experiments. It can have these and other attributes and the resulting $P(x,y,z,t)$ is still a positive real scaler so it can easily be made to satisfy the requirement of a probability variable, that is, $P(x,y,z,t)$ is a positive and real scaler. Thus, the probability of finding an electron in this final choice is

$$P(x,y,z,t)\Delta x\Delta y\Delta z\Delta t = \psi^*(r,t)\psi(r,t)\Delta x\Delta y\Delta z\Delta t$$

To find the differential equation which $\psi(x,y,z,t)$ must satisfy we note from the experimental results listed in Table 141.1 that the solutions or $\psi(x,y,z,t)$ must have several common properties.

(1) $\psi(x,y,z,t)$ must be linear so that its solutions can be superposed (added linearly in appropriate amounts) to produce the interference effect (Davisson-Germer electron diffraction by a crystal) and to produce a wave packet that represents an electron traveling in space whose group (packet) velocity is the particle velocity.

(2) The coefficients of the equation must involve only constants such as $\hbar$, mass m, and charge q, of the material particle, and not the parameters of motion such as momentum, energy, frequency and wave number (p, E, ω, and k). This restriction is necessary in order to allow us to combine the solutions of different p and E to construct more general solutions of an arbitrary experiment.

From these two requirements, we can show that the partial differential equation for electromagnetic and sound waves in classical physics

$$\partial^2\psi/\partial x^2 = A_2\partial^2\psi/\partial t^2,$$

is not a good equation because the coefficient of the differential equation, A_2, is a function of E or p which violates requirement (2). The proof of this violation is left as a homework problem (P142.2). Thus, we next try a linear differential equation which has a first order time derivative.

$$\partial^2\psi/\partial x^2 = A_1\partial\psi/\partial t. \tag{142.1}$$

To determine if the two requirements are satisfied, it is tested by the simple traveling wave solution, $\psi(x,t)=\exp[i(kx-\omega t)]$. Substituting this into (142.1), we found that $A_1=i\omega/k^2=i\hbar E/p^2=i\hbar/2m$ after using the Planck and de Broglie hypotheses $E=\hbar\omega$ and $p=\hbar k$, and $E=p^2/2m=\hbar^2k^2/2m$. This result for A_1 is independent of E and p. Thus, (142.1) is a good equation which satisfies the second criterion. The first criterion was already satisfied at the onset when we picked a linear partial differential equation such as $\partial^2\psi/\partial x^2=A_2\partial^2\psi/\partial t^2$, which does not satisfy property (2), and $\partial^2\psi/\partial x^2=A_1\partial\psi/\partial t$ which does. Substituting $A_1=i\hbar/2m$ into (142.1), we have

$$-(\hbar^2/2m)(\partial^2\psi/\partial x^2) = i\hbar(\partial\psi/\partial t). \tag{142.2}$$

This is the differential equation, known as the **Schrödinger equation**, for a particle of mass m in vacuum or free space which has zero or a spatially constant and time-independent (space-time constant) potential energy.

To obtain a more general differential equation that can describe the motion of a material particle in a region where the potential energy is not a constant, we substitute the plane wave solution $\psi(x,t)=\exp[i(kx-\omega t)]$ into the above equation in order to give us a hint on how to include the potential energy, V(x,y,z). For this purpose, it is more convenient to write the plane wave in terms of the particle momentum and energy using the de Broglie and Planck relations, thus,

$$\psi(x,t) = \exp[i(kx-\omega t)] \tag{142.3A}$$

$$= \exp[i(px-Et)/\hbar]. \tag{142.3B}$$

Using (142.3B), the time derivative term on the right of (142.2) gives

$$i\hbar(\partial\psi/\partial t) = i\hbar(-iE/\hbar)\psi = E\psi \tag{142.4}$$

which suggests that the total energy E can be replaced by the differential operator $+i\hbar(\partial/\partial t)$,

$$E \rightarrow +\ i\hbar(\partial/\partial t). \tag{142.5}$$

Similarly, the space derivative term on the left of (142.2) gives

$$-(\hbar^2/2m)(\partial^2\psi/\partial x^2) = -(\hbar^2/2m)(-p/\hbar)^2\psi$$

$$= (p^2/2m)\psi = T\psi \tag{142.6}$$

which suggests that the kinetic energy T can be replaced by the differential operator $-(\hbar^2/2m)(\partial^2/\partial x^2)$

$$T \rightarrow -\ (\hbar^2/2m)(\partial^2/\partial x^2) \tag{142.7}$$

or the momentum p by the differential operator $-i\hbar(\partial/\partial x)$

$$p \rightarrow -\ i\hbar(\partial/\partial x). \tag{142.8}$$

Substituting the original wavefunction, $\exp[i(kx-\omega t)]$ from (142.3A), into the Schrödinger equation (142.2), we get

$$(\hbar^2k^2/2m)\ \psi(x,t) = \hbar\omega\psi(x,t), \tag{142.9}$$

which reduces to the following form using the de Broglie and Planck relationships, $p=\hbar k$ and $E=\hbar\omega$.

$$(p^2/2m)\ \psi(x,t) = T\psi(x,t) = E\psi(x,t). \tag{142.10}$$

This can also be obtained by substituting the two sets of results just obtained, (142.4) and (142.6), into the Schrödinger equation (142.2).

It is evident that the final equation, (142.10), resembles the classical energy equation for this case of zero potential energy:

$$\text{Kinetic energy} = T = p^2/2m = E = \text{total energy}. \tag{142.11}$$

In a more general problem there is usually a force acting on the material particle. This force is derivable from a spatially varying potential energy, $F=-\partial V(x)/\partial x$.

The result of (142.11) suggests that we add the potential energy to the kinetic energy to give the total energy. Thus,

$$E = T + V$$

and

$$E\psi = (T + V)\psi.$$

We then replace E and T by the operators derived in (142.5) and (142.7). This then gives the generalized one-dimensional Schrödinger equation of a particle with a spatially varying but time-independent potential energy, $V(x)$,

$$i\hbar(\partial\psi/\partial t) = [-(\hbar^2/2m)(\partial^2/\partial x^2) + V(x)]\psi. \qquad (142.12)$$

Although we have not proved that the potential energy term has no operator equivalent and need not be replaced by a differential or integral operator, it is a logical choice since it is derived from a force acting on the material particle. This restriction on $V(x)$ or $V(\mathbf{r})$ is removed when electron spin and relativistic effects are taken into account whose mathematics is given in advanced quantum courses. The assumption of scaler time-independent potential energy can also be extended if there is a time-dependent electromagnetic field like that encountered in the material-light interaction problems such as a p/n junction laser diode, or a time-dependent force from a moving or oscillating ion or ions such as the host ions in the solid lattice. Another example is the use of a time-dependent potential to 'simulate' or to 'model' a potential well in which the electron recombines with a hole in a semiconductor.

From this substitution rule, we can generalize to three space dimensions with a spatially varying potential energy using $E = T + V = p^2/2m + V(\mathbf{r})$ where p^2 now has three spatial (x, y and z) components. The three-dimensional Schrödinger equation of a particle wavefunction, $\psi(\mathbf{r},t) = \psi(x,y,z,t)$, in a potential, $V(\mathbf{r}) = V(x,y,z)$, is then

$$E\psi = (T + V)\psi = (p^2/2m + V)\psi \qquad (142.13)$$

or writing out the operators explicitly, it becomes

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right]\psi + V\psi. \qquad (142.14)$$

The validity of this generalization must again be tested by comparing the theoretical solutions with experimental data. The principal advantage of the above derivation is that it is based on two well-known classical laws, the conservation of energy and Newton's force law, and two quantum postulates, the Planck and the de Broglie's hypotheses. Thus, Schrödinger's equation may be thought of as the mathematical representation of the four postulates and its success depends on experimental verification.

143 Properties and Interpretations of the Wavefunction
(Classical-Quantum Connections)

Five of the more important fundamental properties of the wavefunction are summarized in the following paragraphs.

1. $P(r,t) = \psi^*(r,t)\psi(r,t)$ is the **probability density** (volume density) of finding the particle at position r and at time t.

2. $P(r,t)d^3r = P(r,t)r^2 dr\sin\theta d\theta d\phi = P(x,y,z,t)dxdydz$ is the **probability** of finding the particle in the volume element d^3r, $r^2 dr\sin\theta d\theta d\phi$ or dxdydz, at the position r or (x,y,z) at the time t. The second form uses the spherical coordinate system while the third, the Cartesian coordinate system.

From the probability definition, then,

$$\int_{\text{all space}} P(r,t)d^3r = \iiint \psi^*(x,y,z,t)\psi(x,y,z,t)dxdydz = 1 \qquad (143.1)$$

The physical meaning of the above integral is that if we integrate the position probability density over all space ($x=-\infty$ to $+\infty$, $y=-\infty$ to $+\infty$, and $z=-\infty$ to $+\infty$), we must find the particle, i.e., the probability of finding the particle is unity or we will find the particle for certainty. This is also known as the **normalization condition**, that is, the wavefunction is normalized.

3. The wavefunction ψ and its derivative $\partial\psi/\partial x$, $\partial\psi/\partial y$, $\partial\psi/\partial z$ or its gradient, $\nabla\psi$, must be continuous everywhere. The reason for ψ to be continuous is because it gives the position probability of the particle. It would be physically meaningless to have two values of ψ at one point in space or to have a discontinuity of ψ at a point (x,y,z).

The reason for the slope of the wavefunction or the gradient of the wavefunction, $\nabla\psi$, to be smoothly varying or continuous is that the particle flux density is proportional to the gradient of ψ, $\nabla\psi$. For example, the particle flux (or particle current) in the x-direction is proportional to $\partial\psi/\partial x$ and the particle flux or current cannot be discontinuous physically. The derivation or proof of this result involves a partial integration and a transformation of a volume integral into a surface integral (using Green's Theorem of 3-dimensional calculus). By taking the partial derivative of the probability density integrated over a finite volume, Ω,

$$(\partial/\partial t)\int_{\Omega} P(r,t)d^3r \qquad (143.2)$$

one can get the continuity equation,

$$\partial P(r,t)/\partial t + \nabla\cdot S(r,t) = 0, \qquad (142.3)$$

where Ω is a finite volume. Here S(r,t) is the particle flux density or particle probability current density and it is defined by

$$S(r,t) = (\hbar/2im)[\psi^*\nabla\psi - (\nabla\psi)\psi^*] \qquad (143.4)$$

This result shows that the particle probability current density is proportional to $\partial\psi/\partial x$, $\partial\psi/\partial y$, $\partial\psi/\partial z$, or $\nabla\psi$. Thus, it would be physically meaningless to have a discontinuous change of the slope of wavefunction at a point (x,y,z) in space since that would mean the particle flux or current has two values at that point which is not permissible.

In semiconductor physics, indeed the particle number may be time dependent in contrast to property 2 and the particle flux can be discontinuous in contrast to property 3 both due to the generation, recombination and trapping of the electrons and holes at the trapping centers in a volume element, dxdydz, on a surface or in a thin interfacial layer. These can however be analyzed by extending the above principles to include a time dependent potential energy term which would modify the derived conditions given in properties 2 and 3.

4. As a consequence of property 2, the wavefunction must vanish at large distances, such as a wave packet representing a particle, or a particle bound to a potential well. But for a traveling harmonic wave, such as $\exp(ikx-\omega t)$, it is a constant over all space and at large distance. Thus, for mathematical expediency to avoid the indeterminancy of [(infinitely large number of electrons) + (infinitely large volume of a solid)] while computing the density, a box normalization condition is used. The final solution of the problem would be independent of the box size (as long as it is very large) as expected since it must be physically realistic.

5. Ehrenfest's Theorem
(Correspondence Principle: Quantum-Classical Analogy)

Ehrenfest's theorem deals with the one-to-one analogy between classical and quantum mechanics. It is known as the correspondence principle. We give three examples. For expedient notation we write $\psi\equiv\psi(x,y,z,t)$. The definitions of the expectation value of a function F(x,y,z,t) or an operation 'operator' or 'O' are given by

$$<F> = \iiint_{\text{all space}} \psi^*(x,y,z,t)F(x,y,z,t)\psi(x,y,z,t)d^3r \qquad (143.5)$$

and

$$<O> = \iiint_{\text{all space}} \psi^*(x,y,z,t)O(x,y,z,t)\psi(x,y,z,t)d^3r \qquad (143.6)$$

The classical analogy of quantum mechanical average of the three most important relations in classical mechanics are given in Table 143.1. They can be derived by partial integration. For example, $d\langle x\rangle/dt = \langle p_x\rangle/m$ can be derived by partial integration twice.

Table 143.1
Quantum-Classical Analogy
Ehrenfest's Correspondence Theorem

QUANTUM MECHANICAL AVERAGE		CLASSICAL ANALOGY
$d\langle x\rangle/dt = \langle -i\hbar\partial/\partial x\rangle/m$	$= \langle p_x\rangle/m$	Velocity = Momentum + Mass
$d\langle p_x\rangle/dt = \langle -\partial V/\partial x\rangle$	$= \langle F_x\rangle$	Force = Rate of Momentum Change (Newton's Law)
$d\langle E\rangle/dt = d\langle T+V\rangle/dt$	$= 0$	Conservation of Energy, Power

150 SOLUTIONS OF THE SCHRÖDINGER EQUATION

We will now describe several simple and yet important **steady-state** or stationary solutions of the Schrödinger equation of an electron in idealized one- and three-dimensional potential wells in vacuum. The vacuum solution simplifies the algebra necessary to solve the Schrödinger equation in order to serve as simplified model examples of electron transport phenomena that can occur also in insulators, semiconductors and metals. The solutions of the steady-state electron wavefunction in an isolated hydrogen atom in vacuum are given as the last example (section 156) since they serve as the base for solutions in the multi-electron atoms given in the following section (160) and in the many-electrons and many-atomic nuclei in a solid (sections 17n).

The equation which the stationary solutions must satisfy can be obtained from the time-dependent Schrödinger equation, (142.12) or (142.14), by means of the method of separation of variables in differential equation theory. Consider the one-dimensional space for simplicity, then the general time-dependent solution can be separated into a product of a space function and a time function, $\psi(x,t)=\psi(x)T(t)$, based on the theory of separation of variables. Substituting this assumed product solution in (142.12), and dividing both sides of the equal sign by $\psi(x)T(x)$, then

$$\frac{i\hbar[\partial T(t)/\partial t]}{T(t)} = \frac{[-(\hbar^2/2m)(\partial^2/\partial x^2) + V(x)]\psi(x)}{\psi(x)} . \qquad (150.0)$$

Since T(t) and $\psi(x)$ are arbitrary, the above must be equal to a constant which is denoted by E. Thus, the time-dependent Schrödinger equation is now separated into two equations, a time and a space equation.

$$i\hbar[\partial T(t)/\partial t] = ET(t) \tag{150.0A}$$

$$[-(\hbar^2/2m)(\partial^2/\partial x^2) + V(x)]\psi(x) = E\psi(x). \tag{150.0B}$$

The time dependent equation, (150.0A), can be integrated to give

$$T(t) = A\cdot\exp(Et/i\hbar) \equiv A\cdot\exp(-i\omega t) \tag{150.0C}$$

where Planck's condition $E=\hbar\omega$ is used and A is a constant.

The space equation, (150.0B), is the stationary, steady-state, or time-independent Schrödinger equation. We shall describe several space solutions after their general properties are discussed in the following paragraphs. The total solution is then given by the product

$$\psi(x,t) = \psi(x)T(t) = \psi(x)\exp(Et/i\hbar) = \psi(x)\exp(-i\omega t). \tag{150.0D}$$

General Properties

To sketch an electron wavefunction in a given spatial variation of the potential energy, four helpful rules are noted based on physical reasoning and the correspondence principle. These are listed and discussed in the following paragraphs.

(1) $\psi(x)$ is continuous. (150.1)

(2) $d\psi/dx$ is continuous. (150.2)

(3) The spatial variation of the wave length of oscillation or of the electron wavefunction, λ, is given by the de Broglie relationship, $\lambda=h/p$, and the classical energy-momentum relationship, $E=KE+PE=p^2/2m+V$.

When the kinetic energy is positive or $T=E-V > 0$, the wavelength of the electron wave is then given by

$$\lambda = h/\sqrt{[2m(E-V)]}. \tag{150.3}$$

The solution is sinusoidal in space if E-V=constant, or aperiodic if E-V=f(x) varies with position.

In the classically forbidden region, $E-V=T<0$, then the wave length is imaginary and physically meaningless. The wave number is also imaginary and given by $k = 2\pi/\lambda = \hbar^{-1}\sqrt{[2m(E-V)]} = i\hbar^{-1}\sqrt{[2m(V-E)]}$. Thus, instead of the oscillatory plane wave solution $\exp(\pm ikx)$, the solutions are exponentially rising or decaying in space given by $\exp(\pm|k|x)$.

Note that the wavelength of the probability of finding an electron, $P(x,t)=\psi^*\psi$, is half that of ψ, for example, if $\psi=\sin(kx)$ then $P(x,t) = \sin^2(kx) = [1-\cos(2kx)]/2$ so that the wave length of P is $2k=2\pi/\lambda_2$ or $\lambda_2=\pi/k=\lambda/2$.

(4) The amplitude of P(x,t) is proportional to the time the particle spent in a region or inversely proportional to the particle velocity or particle momentum. Thus, the amplitude can be estimated by

$$\text{AMPLITUDE} = A/v = B/p = C/\surd(E-V) \qquad (150.4)$$

in the classically allowed regions of space where $E-V > 0$. This rule is good only when the energy is large and the quantum solution approaches the classical solution. Nevertheless, it gives a good classical sense of the significance of a quantum solution. An example is the high-quantum-number solutions of an electron in the square well potential. Another example is the high-quantum-number solution of an atom oscillating in a harmonic oscillator potential well, $V(x) \propto -x^2$.

We shall next describe the techniques of solving the Schrödinger equation and solutions of several physical problems in vacuum which have a counter part in solid state and semiconductor device physics. They are used by pioneers who invented or discovered some device phenomena and solved the zeroth and first order theory. These one-electron solutions in vacuum or free space can be adapted to the many-electron and many-atom solid after a one-electron transformation is applied to the many-electron system of a solid. The selected one-electron solutions in free space are well-known examples derived in the introductory lectures on atomic and quantum physics of matter taught in sophomore general physics and junior modern physics courses. We include them here to illustrate solution techniques, and also for ease of reference and uniformity of notation when they are used in those semiconductor devices whose operation is dominantly controlled by one of these quantum or wave phenomena.

151 Reflection of Electron at a Potential Step

This was the model used by Shockley and Bardeen in 1951 to describe the probability of scattering of electrons and holes in a semiconductor by the random vibration of the atoms and ions of the crystal. It was known as the **deformation potential model** and it was used to calculate the mobility and drift velocity or drift current of electrons and holes. It has been a popular model and given in intermediate textbooks on semiconductor physics and solid state device theory. Recently, it is also used as a model to describe the electrons passing through an interface between two materials that has a interfacial potential discontinuity. Examples are the semiconductor heterojunctions studied in the recent research and development focus on advanced compound semiconductors for ultra-high-speed transistors, the metal/semiconductor or Schottky barrier diode, and the oxide/silicon interface in the metal-oxide-semiconductor field-effect transistor which is the mainstay of integrated circuit technology of the present and future generations.

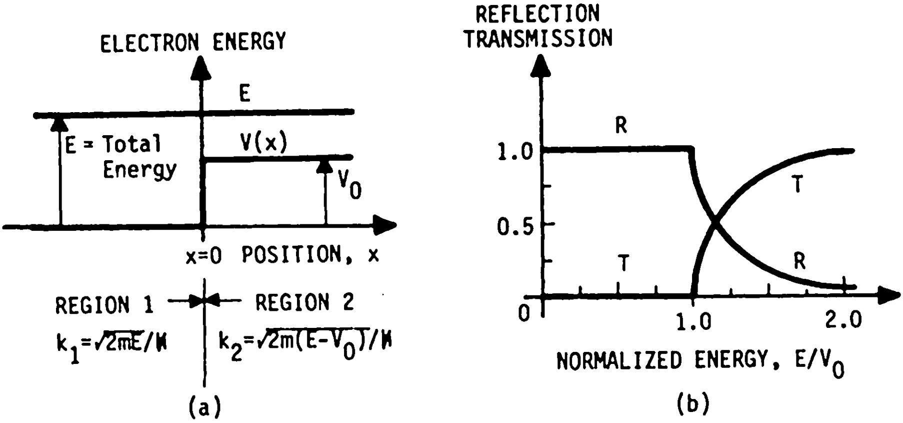


Fig.151.1 (a) The electron potential energy diagram V(x) where the total energy is plotted at a function of position, x. (b) The reflection and transmission coefficients obtained from the solutions as a function of the electron energy normalized to the size of the potential step.

We shall write down the algebraic steps used to get to the solution without wordy elaboration. The solution on the left, x<0, or the side-1, denoted as REGION 1, is given generally by two waves travelling in opposite directions. [Recall, $\psi(x,t)=\psi(x)\exp(-i\omega t)$.]

$$\psi_1(x) = A_1\exp(ik_1x)+B_1\exp(-ik_1x) \qquad (151.1)$$

since the potential step reflects part of the electron wave traveling to the right and reflection is represented by the $B_1\exp(-ik_1x)$ term. The general solution on the right, x>0, or side-2, REGION 2, is given by

$$\psi_2(x) = A_2\exp(ik_2x) \qquad (151.2)$$

since the transmitted wave travels to the right and there is no reflected wave traveling towards the left in this region. Generally, we would have $A_2\exp(ik_2x)+B_2\exp(-ik_2x)$ but we have $B_2=0$ in this case since there is no wall at $x\to+\infty$ to reflect the wave.

Matching the two wavefunctions and their derivatives at the boundary, $x=0$, we get the following two equations for the coefficients. From $\psi_1(0)=\psi_2(0)$, we get

$$A_1 + B_1 = A_2. \tag{151.3}$$

From $d\psi_1(x)/dx = d\psi_2(x)/dx$ at $x=0$, we get

$$k_1(A_1 - B_1) = k_2A_2. \tag{151.4}$$

Solving these two boundary equations, two of the three unknowns (A_1, B_1 and A_2) can be eliminated. Retain A_1 and express B_1 and A_2 in term of A_1, then, the results are

$$B_1/A_1 = (k_1-k_2)/(k_1+k_2) \tag{151.5}$$

and

$$A_2/A_1 = 2k_2/(k_1+k_2). \tag{151.6}$$

Thus, the wavefunctions are:

$$\psi_1(x) = A_1\left[\exp(ik_1x) + \frac{k_1-k_2}{k_1+k_2}\exp(-ik_1x)\right] \quad \text{for } x<0 \tag{151.7}$$

and

$$\psi_2(x) = A_1\left[\frac{2k_1}{k_1+k_2}\exp(ik_2x)\right]. \tag{151.8}$$

A_1 can be obtained from normalization in a box of length 2L extending from -L to +L and L is made to approach infinity for a problem containing all space. But, we only need to know the reflection and transmission coefficients or their ratio which cannot depend on how we normalize the wavefunction. The reflection coefficient is defined as $|v_1B_1/v_1A_1|^2$ since the particle flux is proportional to its velocity v times its density, $\psi^*\psi$. The incident particle flux or current is $v_1\psi_1{}^*\psi_1=v_1|A_1|^2$. The velocity can be obtained from $p=mv=\hbar k$ for a traveling harmonic wave such as exp(ikx). Using these, then, the reflection coefficient is

$$R = |v_1B_1/v_1A_1|^2 = [(k_1-k_2)/(k_1+k_2)]^2. \tag{151.9}$$

Similarly, the transmission coefficient is defined by $|v_2A_2/v_1A_1|$ and given by

$$T = |v_2A_2/v_1A_1|^2 = (k_2/k_1)[2k_2/(k_1+k_2)]^2 = 4k_1k_2/(k_1+k_2)^2. \tag{151.10}$$

$E=KE+PE=(\hbar^2k^2/2m)+V$ is used to get $k_1=\surd(2mE)/\hbar$ and $k_2=\surd[2m(E-V_0)]/\hbar$. Note that $R+T=1$ as expected. A sketch of T and R vs the electron energy E is given in Fig.151.1(b). It can be shown that half of the incident electron is transmitted ($T=R=0.5$) when $k_1/k_2=3+2\surd2=5.828$ or $E/V_0=1.03033$. It can also be shown that the solution $k_1/k_2=3-2\surd2=0.1715729$ is not permitted.

152 Resonance Scattering by a Square Potential Well

It has been known that a light filter will pass light if the thickness of the layer in the filter is a multiple of the half-wavelength of light. The square well potential, either attractive or repulsive as shown in Figs.152.1(a) and (b) respectively, has a similar filtering property for the electron wave. The latest compound semiconductor heterojunction transistor using a thin quantum well for its base layer makes use of this resonance transmission. A large electron transmission or current transport across such a thin base layer could be attained in principle if the base layer is made half- or one-wavelength thick. Then, it serves as a selective filter to transmit electrons with a certain kinetic energy or wave length. The thin layers are grown by the molecular beam epitaxy (MBE) technique. The thickness of these layers or wells is only a few atomic distances or contains only a few atomic layers. Devices with many layers and multiple negative resistance regions in the I-V curves have been fabricated in the laboratory which have transition times of a few picoseconds between the negative resistance states. The structure is in a highly thermodynamic nonequilibrium state because of the atomically thin layers of different chemical compositions, suggesting that its endurance or operating life must be severely limited by interlayer atomic diffusion. However, its operating life and reliability have not been addressed by the researchers.

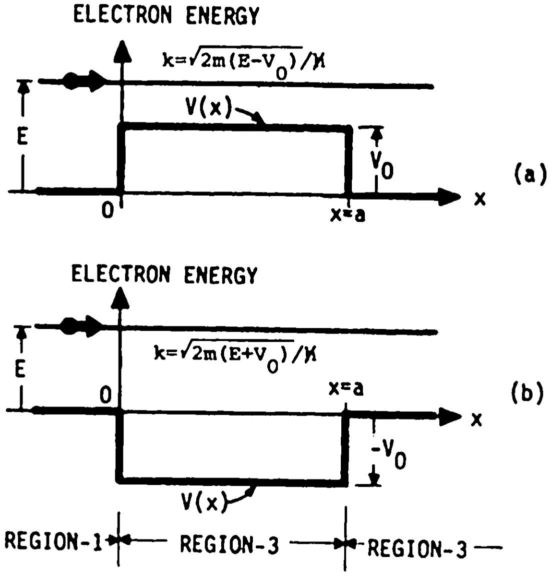


Fig.152.1 The repulsive (a) and attractive (b) potential wells for the analysis of resonance scattering of electron waves.

The general solution of this three-region potential barrier problem can be obtained in the same way as the two-region problem given in section 151. The present problem requires much more algebra since there are three regions and three wavefunctions, one in each region, and since the wavefunctions and their derivatives must be matched at the two boundaries, $x=0$ and $x=a$. However, the solutions at the resonance energies are particularly easy to derive since at resonance, the incident electron is completely transmitted and there is only a phase shift and no reflection. We will work out this special case and leave the general solution as a problem for advanced students. The potential and total energies are shown in Figs.152.1(a) and (b) whose coordinate gives the simplest arithmetic. It is not necessary to translate the coordinate to make the potential well symmetrical with respect to $x=0$ since the wavefunctions are not symmetrical anyway. The solutions in the three regions are

$$\psi_1(x) = \exp(ik_1x) \qquad x \leq 0 \qquad (152.1)$$

$$\psi_2(x) = A\exp(ikx) + B\exp(-ikx) \qquad 0 \leq x \leq a \qquad (152.2)$$

and

$$\psi_3(x) = \exp[i(k_1x - n\pi)] \qquad x \geq a \qquad (152.3)$$

The factor $n\pi$ in the transmitted wave in region 3 represents the total phase shift experienced by the wave while passing through the middle or barrier layer (region 2). It arises from the condition that at resonance the thickness of the barrier layer is an integer number of half-wavelength of the electron wave, i.e., $a=(\lambda/2)n$ where $n=1,2,3,4,\ldots$.

To evaluate A and B, ψ_1 and ψ_2 and their derivatives are matched at $x=0$ which give

$$A + B = 1 \qquad (152.4)$$

and

$$A - B = k_1/k \qquad (152.5)$$

so that

$$A = (k + k_1)/2k \qquad (152.6)$$

and

$$B = (k - k_1)/2k. \qquad (152.7)$$

Then, the wavefunction in the barrier layer (region II) is

$$\psi_2(x) = \cos(kx) + i(k_1/k)\sin(kx). \qquad (152.8)$$

Properties of the solutions are described by the problems 152.1 to 152.6 listed at the end of chapter 1.

153 Tunneling through a Square Potential Barrier

When the total electron energy in Fig.152.1(a) is less than the repulsive potential barrier height, $0<E<V_0$, classical mechanics states that the electron should be completely reflected. However, quantum mechanically, there is a non-zero probability that the electron will be transmitted through the classically forbidden layer, $0<x<a$, where the potential energy is larger than the kinetic energy. This is known as **tunneling**. The electron energy diagram is given in Fig.153.1(a). The electron position probability density, to be obtained, is given in Figs.153.1(b) and (c).

This is a model for the quantum mechanical tunneling phenomenon that occurs in metal/insulator/metal, metal/insulator/semiconductor, and semiconductor/insulator/semiconductor diodes. Significant current can pass through such a diode if the insulator layer is very thin.

To make the mathematical problem simple, we assume that the electron wave travels from the left to the right and we move the coordinate origin to the exit or right edge of the potential barrier as shown in Fig.153.1(c) and label the new horizontal axis by 'y'. When the electron wave hits the left wall at $y=-a$, part of it is reflected and move backwards in the negative y direction. The remainder penetrates into the classically forbidden potential barrier where it has a decaying amplitude since the kinetic energy is negative or the wave number k is imaginary as illustrated in Fig.153.1(c).

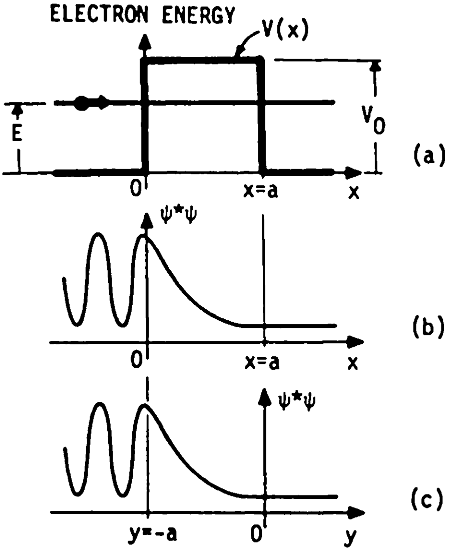


Fig.153.1 Tunneling through a square potential barrier. (a) The potential energy diagram of an electron. (b) The probability amplitude as a function of position. (c) Shifting the origin of the coordinate to simplify the algebra of the solution.

Again, this is a two-boundary problem which requires matching the solutions of the three layers at the two boundaries. Although the complete solution is complicated, a simple solution can be quickly obtained for the barrier region in terms of the transmitted wave, since we know that the transmitted wave is given by

$$\psi_3(x)=A_3\exp(ikx) \qquad \text{Fig.153.1(b)}$$

or

$$\psi_3(y)=A_3\exp(iky) \qquad \text{Fig.153.1(c)} \tag{153.1}$$

where

$$k=\hbar^{-1}/\sqrt{(2mE)}. \tag{153.2}$$

For simplicity without losing any generality, we let $A_3=1$. The wavefunction inside the barrier must have two terms since the potential step at $x=a$ or $y=0$ will reflect part of the wave. Thus,

$$\psi_2(x) = A_2\exp(\alpha x) + B_2\exp(-\alpha x)$$

or

$$\psi_2(y) = C_2\exp(\alpha y) + D_2\exp(-\alpha y) \tag{153.3}$$

where

$$\alpha = \sqrt{[2m(V_0-E)]}/\hbar. \tag{153.4}$$

The constants, A_2 and B_2, are obtained by matching $\psi_2(x)$ and $\psi_3(x)$ and their derivatives at $x=a$ or $y=0$. The algebra is simpler using the y-coordinate. The match gives

$$C_2 + D_2 = 1$$

and

$$C_2 - D_2 = ik/\alpha \tag{153.5}$$

so that

$$C_2 = (1 + ik/\alpha)/2$$

and

$$D_2 = (1 - ik/\alpha)/2. \tag{153.6}$$

The wavefunction in the barrier region is then given by

$$\psi_2(y) = \cosh(\alpha y) + i(k/\alpha)\sinh(\alpha y). \tag{153.7}$$

The probability density functions are:

$$P_3(y) = 1 \tag{153.8}$$

and

$$P_2(y) = \psi_2{}^*\psi_2 = \cosh^2(\alpha y) + (k/\alpha)^2\sinh^2(\alpha y). \tag{153.9}$$

They are those given in Fig.153.1(c). The tunneling probability through a square potential barrier of thickness a is then approximately given by

$$T = P_3(y=0)/P_1(y=-a) \tag{153.10}$$

$$\approx P_3(y=0)/P_2(y=-a) \qquad \text{(153.10A}$$

$$= 1/[\cosh^2(\alpha a)+(k/\alpha)^2\sinh^2(\alpha a)] \approx 4\exp(-2\alpha a) \tag{153.11}$$

which assumes a high potential barrier, $V_0>>E$, such that $k/\alpha = \sqrt{E/(V_0-E)} << 1$, and a thick potential barrier such that $\alpha a = (a/\hbar)\sqrt{[2m(V_0-E)]} >> 1$.

154 Tunneling through a Triangular Potential Barrier

In metal/oxide-insulator/semiconductor or MOS diodes and semiconductor p/n junction diodes, tunneling usually becomes important only when there is a high electric field. In MOS diodes, the potential barrier is nearly triangular like that shown in Figs.154.1(a) and (b).

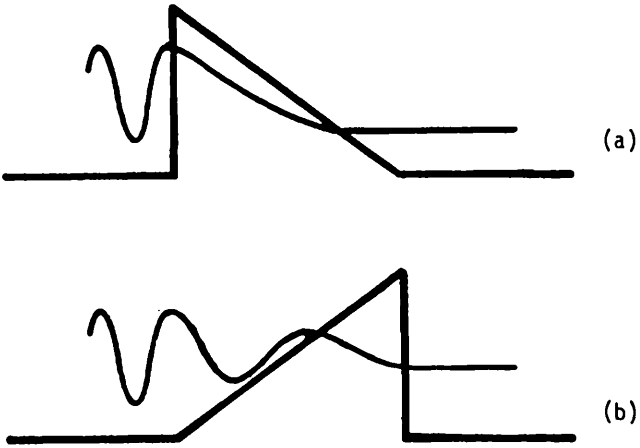


Fig.154.1 Tunneling of electron through a triangular potential barrier. (a) Incident from the abrupt wall. (b) Incident from the inclined wall.

Since the slope is proportional to the electric field, obviously, high electric field enhances tunneling since it reduces the barrier thickness, although the barrier height is unchanged. The mathematics of solving the Schrödinger equation in a spatially constant or varying electric field is complicated by the position dependence of the potential energy function in the barrier region. However, when the barrier is high and thick, the tunneling probability can be evaluated by a very simple formulae, known as the **tunneling integral** which is given by,

$$T = \exp\left[-2\int\alpha(x)dx\right]_{\text{integrate over forbidden path.}} \qquad (154.1)$$

This simple formula is suggested by the asymptotic solution, $\exp(-2\alpha a)$, of the rectangular potential barrier given by (153.11). When the potential energy varies spatially, the attenuation constant, α, is no longer constant, so one would expect the exponent, αa, to be replaced by an integral over the barrier region where α, defined by (153.4), $\alpha(x)=\hbar^{-1}\surd\{2m[V(x)-E]\}$, is positive or $V(x)>E$, that is, the kinetic energy is negative, $T=E-V(x)<0$. Thus, the integral is taken over the classically **forbidden path**. For the MOS diode with the electron potential energy diagrams given by Figs.154.1(a) and (b), the tunneling probabilities are

$$T_a = \exp\{-[4(2m)^{1/2}/(3\hbar qF)][(V_0-E)^{3/2}]\} \qquad \text{(abrupt entrance wall)} \quad (154.1A)$$

$$T_b = \exp\{-[4(2m)^{1/2}/(3\hbar qF)][(V_0-E)^{3/2}]\} \qquad \text{(inclined entrance wall)} \quad (154.1B)$$

where $\alpha=\hbar^{-1}\surd[2m(V_0-E\pm qFx)]$ and F is the magnitude of the electrical field in V/cm. The integration is carried through the forbidden path, $x = 0$ to $x = \pm x_T = \pm(V_0-E)/qF$, where + is for incidence onto the abrupt wall [Fig.154.1(a)] and - for incidence onto the inclined wall [Fig.154.1(b)]. + also corresponds to a positive voltage applied to the right (+x) terminal of the diode relative to the left (-x) terminal, while - corresponds to a negative applied voltage to the right terminal to give the electric field. The tunneling probability is identical for the two incident directions in this simple approximation. From the potential energy diagrams shown in Figs.154.1(a) and (b), one would expect a difference. One obvious difference is in the tunneling rate which is the tunneling probability times the incident rate and the incident rate is the electron velocity times the electron density. It is obvious that the incident velocity is smaller in the inclined wall case because the incline slows down the electron as it approaches the inclined edge of the wall. Hence incident onto the inclined wall would have a small tunneling rate and current.

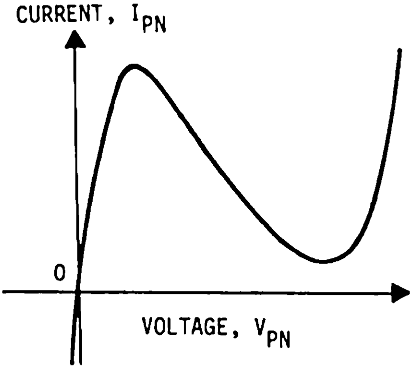


Fig.154.2 The current voltage characteristic of a narrow semiconductor p/n junction or Esaki diode which shows a negative resistance due to quantum mechanical tunneling.

Tunneling current in the MOS diode is increasingly important in present and future state-of-the-art silicon transistors and integrated circuits as the transistor dimension and the oxide-insulator thickness decreases. However, the first direct and unambiguous experimental demonstration and theoretical interpretation of tunneling in semiconductor devices was by Esaki (IBM) while still at Sony in 1957 on a Ge p/n junction diode. When the p/n junction is very thin, a negative conductance current-voltage characteristic, shown in Fig.154.2, was observed. The increasing and then decreasing current with increasing applied diode voltage arises from the increasing and then decreasing density of energy levels to which the electrons can tunnel as the applied voltage, V_{PN}, is increased from zero.

155 Bound States in an Attractive Square Potential Well

In real-world physical problems, whether in vacuum or in solids, bound electron states are solutions of the Schrödinger equation **localized in all three dimensions**. They are centered around an isolated three-dimensional attractive potential well, such as the electron attracted to a proton by the Coulomb force. The object creating the well is known as a **center** or an **electron trap**. There is a very important difference between the one-dimensional (1-d) and three-dimensional (3-d) bound state solutions: one-dimensional bound state solutions exist for any well size and depth; three-dimensional bound state solution does not exist until the well size and depth are sufficiently large. This difference obviously has very important consequences on the conduction of electricity by electrons and holes. However, we shall only describe the steps of solving for the one-dimensional bound states in an attractive potential well in free space in order to illustrate the mathematical technique and the basic physics. The 1-d problem is also a useful model for extending the fundamentals to the two- and three-dimensional wells in vacuum and in semiconductors and solids which have many valence electrons that screen the potential well. Bound electron states are created around localized atomic defects or isolated impurities in crystalline solids since the defects and impurities disrupt the periodic potential, seen by the electron, arising from the ionic cores of the host atoms that sit on the crystal lattice. When a bound electron (or hole) solution of the Schrödinger exists, the defect or impurity is known as an **electron** (or **hole**) **trap**. A conduction electron in a transistor can be trapped at an electron trap. A conduction hole can then recombine with the trapped electron. When the recombination energy is dissipated as heat, it is known as **thermal recombination**. Thus, the conductivity and signal carrying capacity of the conduction electrons and holes are reduced by the presence of traps. Thermal recombination is a fundamental mechanism that limits the current amplification of bipolar transistors. However, if the recombination energy is carried away by the emission of a photon, then we have a useful light emitting photonic device which converts electricity into light. The trapped electron may also be emitted (released, detrapped or ejected) from the trap by the absorption of a photon before it recombines with a hole. This would be useful as a photon detector since it converts light into a conduction electron. A sample of photo solid-state and semiconductor devices includes the phosphor screen of the monochromatic and color TV and computer display tubes, the electroluminescence panels, the semiconductor p/n junction light emitting diodes of different colors, and the photoconductor used in xerography. Thus, the bound state problem of a simple one-dimensional attractive square potential well is a valuable model for understanding the existence and the properties of bound states in real crystalline and even noncrystalline semiconductors and solids. In chapter 3, the effect of bound states on the properties of semiconductor will be described. Its applications to devices are described through the device chapters starting from chapter 4.

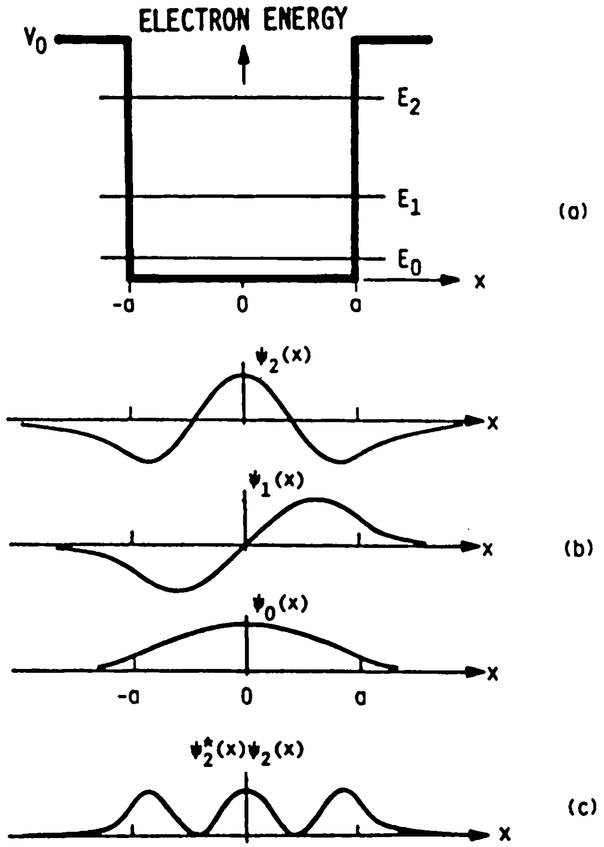


Fig.155.1 Bound state solutions of the attractive one-dimensional square potential well. (a) The well and three lowest energy levels, $n=0$, 1, and 2. (b) The three lowest-energy wavefunctions, ψ_1, ψ_2 and ψ_3. (c) Position probability density, $\psi_2{}^*\psi_2$.

The bound state solutions of the one-dimensional attractive potential well can be readily obtained using the method of the scattering problem described in Fig.152.1(b). The energy of the electron is now negative, i.e., below the top of the square-potential-well shown in Fig.152.1(b). In contrast to the simple solution obtained for the scattering problem of Fig.152.1(b), in order to get the bound state solutions we cannot avoid the tedious algebra arising from the two well boundaries and from three solutions, one each in the three regions. But, one can simplify the algebra tremendously by judicial choice of the coordinate system based on simple physical consideration of the type of solutions one expects to get. The choice is to place the well symmetrically with respect to the coordinate origin. Thus, the attractive square well shown in Fig.152.1(b) is now shifted by a/2 towards the negative x direction as shown in Fig.155.1(a) and the bottom of the well is selected as the reference or zero energy.

An additional simplification of notation and algebra is gotten by making the well thickness 2a instead of 'a' to avoid carrying the factor 1/2 in the algebra. With these choices, the solutions can be written down with a minimum number of undetermined coefficients since they must be either symmetrical or antisymmetrical with respect to x and they must decay exponentially outside of the well. Thus, in regions 1, 3, and 2

$$\psi_1(x) = A_1\exp(+\alpha x) \qquad x < -a \qquad (155.1)$$

$$\psi_3(x) = \pm A_1\exp(-\alpha x) \qquad x > +a \qquad (155.2)$$

$$\psi_2(x) = A_2\sin(+kx) \quad - \text{ sign in } \psi_3 \quad -a > x > +a \quad \text{Asymmetrical} \qquad (155.3)$$

$$\psi_2(x) = A_2\cos(+kx) \quad + \text{ sign in } \psi_3 \quad -a > x > +a \quad \text{Symmetrical} \qquad (155.4)$$

where $\alpha = \surd[2m(V_0-E)]/\hbar$ and $k_2 = \surd[2mE]/\hbar$. The coefficients A_1 and A_2 can be obtained by matching both ψ and $d\psi/dx$ at $x = +a$ or $x = -a$ for the symmetrical as well as the asymmetrical solutions. We skip the algebra. As an elementary exercise, the students can work out the trivial solutions of the infinite well, $V_0 = \infty$, starting from (155.1) to (155.4). As an advanced exercise involving only longer algebra, the students can verify the finite well ($V_0 \neq \infty$) solutions given below. From matching the three wavefunctions and their derivatives at the two boundaries, the following transcendental equations are obtained from which the one-dimensional bound state solutions for the energy can be computed.

$$(ka)\tan(ka) = \alpha a \qquad \text{Symmetrical} \qquad (155.5)$$

and

$$(ka)\text{ctn}(ka) = -\alpha a \qquad \text{Antisymmetrical.} \qquad (155.6)$$

These can be solved numerically and simultaneously with the following equation from the definitions of α and k,

$$(\alpha a)^2 + (ka)^2 = 2mV_0a^2/\hbar^2. \qquad (155.7)$$

This shows that one-dimensional bound state exists even as $V_0 \to 0$ and $a \to 0$. In contrast, for the three-dimensional spherical potential well which is the real-world situation, there is no bound state unless $V_0a^2 > \pi^2\hbar^2/8m$. There is one 3-d bound state if $\pi^2\hbar^2/8m < V_0a^2 < 9\pi^2\hbar^2/8m$. The first of another series of 3-d bound states appear when $4\pi^2\hbar^2/8m \leq V_0a^2 \leq 16\pi^2\hbar^2/8m$. The difference between the existence condition of bound state in the 1-d and 3-d cases indicates the complexity and uniqueness in a real-world situation. It has important consequences on the magnitude of electrical conductivity in 1-d, 2-d, and 3-d conductors.

The three lowest energy wavefunctions and probability amplitudes, $\psi_n^*(x)\psi_n(x)$, are given in Figs.155.1(b) and (c). They are labeled by the integer n=0,1,2. n is selected such that n=0 is the ground state, n=1 is the first excited state, etc. Furthermore, the wavefunction is symmetric or even with respect to x when n=even, and antisymmetric or odd when n=odd. Note that n is also the number of nodes or zeros of $\psi_n(x)$.

156 The Hydrogen Atom

The solutions of the electron wavefunction in the potential well of an array of positively charged nuclei have provided the **physical basis for using the energy band and valence bond models** to describe electrical conduction by the valence electrons and holes in metals, semiconductors, and insulators. For this reason, the wavefunctions and allowed energies of an electron in a hydrogen atom are studied in this section since it has the simplest nucleus, a proton. Although hydrogen is the simplest atom, the mathematics to obtain the solution is tedious since it is a three-dimensional problem. Only the results and their physics base will be described. The tedious algebra of derivation will only be sketched. The detailed algebra of the derivation can be found in all the introductory quantum mechanics textbooks, such as Eisberg and Resnik referred to in Table 141.1.

The potential energy of the electron in the hydrogen atom arises from the attractive electrostatic force between the negatively charged electron and the positively charged proton known as the Coulomb force. The potential energy is given by $V(r) = -q^2/4\pi\varepsilon_0 r$ where r is the distance between the proton and the electron. The electron and the proton are treated as point charges whose radii are small compared with the interparticle distance, r, which has an average value of about 0.5 to 1 Angstrom (1 Angstrom$=$1A$=10^{-8}$cm). The particle radii are defined by specific experiments or are experiment-dependent. For example, an electromagnetic radius was defined for an electron, and the radius of a proton or neutron was obtained in scattering experiments and they were all about 1 Fermi or 10^{-13}cm. This is much smaller than the interparticle distance of 1A or 10^{-8} cm in solids, validating the point charge model.

The solutions have the simplest appearance if the proton is located at the coordinate origin, $r=0$. Then r is the position vector of the electron. This potential energy is spherically symmetric, i.e., independent of angular positions. Hence it is best to find the solutions of the Schrödinger equation in the spherical coordinate (r,θ,ϕ) shown in Fig.156.1 on the following page. Using the geographical convention for the coordinate system of the earth, z is the polar axis, θ is the azimuthal angle (north-south latitude), and ϕ is the polar angle (east-west longitude). In the spherical coordinate system, the kinetic energy operator

$$-(\hbar^2/2m)(\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2)$$

is expressed in the variables (r,θ,ϕ) of the spherical coordinate and the stationary Schrödinger equation becomes

$$-(\hbar^2/2m)\nabla^2\psi(r,\theta,\phi) + V(r)\psi(r,\theta,\phi) = E\psi(r,\theta,\phi) \qquad (156.1)$$

where $\hbar=h/2\pi$ is the Planck constant, m is the rest mass of the electron (assuming the proton is stationary), E is the total energy which is to be determined by the solution. ∇^2 is the Laplacian operator defined by

$$\nabla^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\left[r^2\frac{\partial}{\partial r}\right] + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2} + \frac{1}{r^2\sin^2\theta}\frac{\partial}{\partial\theta}\left[\sin\theta\frac{\partial}{\partial\theta}\right] . \qquad (156.2)$$

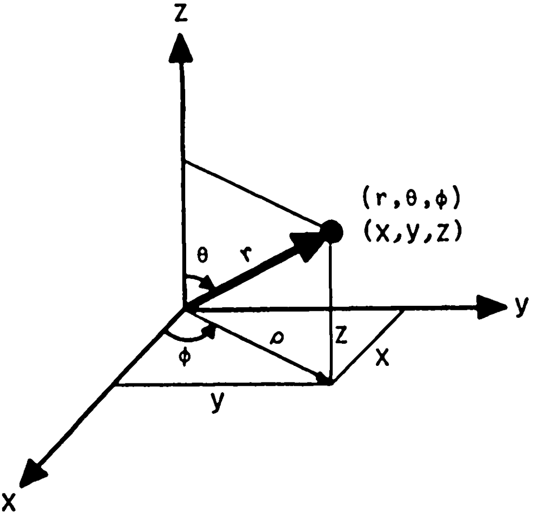


Fig.156.1 The spherical (r,θ,φ) and corresponding Cartesian (x,y,z) coordinate systems used in the solution of the electron wavefunction in a hydrogen atom.

The method of separation of variables in the theory of partial differential equations is then used to solve the Schrödinger equation (156.1). This method assumes that the solutions of the electron wavefunction can be written as a product of three parts:

$$\psi(r,\theta,\phi) = R(r)\Theta(\theta)\Phi(\phi). \qquad (156.3)$$

R(r) is known as the **radial wavefunction**, $\Theta(\theta)$, the **angular** or **azimuth wavefunction**, and $\Phi(\phi)$, the **polar wavefunction**. The product of the azimuth and polar wavefunctions is known as the **spherical harmonics** defined by $Y(\theta,\phi)=\Theta(\theta)\Phi(\phi)$. Each of these parts and the wavefunction itself are all normalized so that the integral over all space of the square of their amplitude is equal to unity to conform with the probability interpretation.

Substituting this three-part product-solution into the Schrödinger Equation (SE) and dividing the SE by this product, we get three ordinary differential equations for the three parts which are given and explained in the following paragraphs.

The polar wave equation is

$$\frac{d^2\Phi}{d\phi^2} = -m_L{}^2\Phi \qquad m_L = -L,\ldots\ldots,-1,\ 0,\ +1,\ \ldots\ldots,+L. \tag{156.4}$$

where the integer m_L is obtained by using the periodic boundary condition, $\Phi(0)=\Phi(2\pi)$. It is known as the **magnetic quantum number** since it measures the importance of the magnetic field on the solution and determines the electron energy if the hydrogen atom is immersed in a magnetic field.

The azimuthal wave equation is

$$-\frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{d\Theta}{d\theta}\right] + \frac{m^2\Theta}{\sin^2\theta} = L(L+1)\Theta \quad L=0,1,2,3,4\ldots.(n-1) \tag{156.5}$$

where $m=|m_L|$. L is known as the **azimuthal, orbital** or **angular momentum quantum number**. (Normally, a scripted lower-case L, *l*, is used. Due to word-processor limitation, a capitalized L is used here.) n is known as the **principal quantum number**. This equation is known as the **Legendre Equation** and its solutions, Θ, are the **associated Legendre functions**, denoted by $P_L{}^m(\cos\theta)$.

The radial wave equation is

$$\frac{1}{r^2}\frac{d}{dr}\left[r^2\frac{dR}{dr}\right] + \frac{2m}{\hbar^2}\left[E - V(r)\right]R = L(L+1)\frac{R}{r^2} \tag{156.6}$$

where $\hbar=h/2\pi$. For a point charge with $+Zq$ charges ($Z=1$ for proton in the hydrogen problem), $V(r)=-Zq^2/4\pi\varepsilon_o r$. The solutions can be written as a product given by

$$R_{nL}(r)=[\exp(-Zr/na_0)]\cdot[Zr/na_0]^L\cdot G_{nL}(Zr/a_0).$$

$G_{nL}(Zr/a_0)$ is a polynomial of Zr/a_0 and is known as the **Laguerre polynomials**. n is the **principal quantum number** given by $n=1,2,3,\ldots.,\infty$. a_0 is the normalization radius defined by

$$a_0=4\pi\varepsilon_o\hbar^2/mq^2. \tag{156.6A}$$

The allowed solutions have electron energy given by

$$E_n = - Z^2q^4m/[2\hbar^2(4\pi\varepsilon_o)^2n^2] = -13.605(Z^2/n^2) \text{ eV} \qquad (156.6B)$$

which is identical to the result of the simple Bohr atom model given by (141.8) when $Z=1$ is used.

The key results are (i) the **allowed electron energies are quantized or discrete**, and (ii) the wavefunctions **are localized** at or **bound to** the nucleus, that is, it decays exponentially from the positively charged nucleus. This bound or localization property is also reflected by the negative total energy relative to the vacuum level (the top of the potential hill). Thus, the electron is said to be **bound** or **trapped** to the **potential well** from the positively charged proton or the positive point charge, $+Zq$.

In summary, the following nomenclatures are introduced and defined: n = **principal quantum number**, L = **azimuthal, orbital** or **angular momentum quantum number**, m_L = **magnetic quantum number**. The electron wavefunctions around a proton is also known as the **hydrogen orbitals** to chemists. For a given L they are designated by a letter in atomic spectroscopy and chemistry, for example, s for L=0, p for L=1, d for L=2, ...

The normalized wavefunctions labeled by the spectroscopic or chemistry subscripts are given in Table 156.1, for example, $\psi_{100}(r,\theta,\phi)\equiv\psi_{1s}$. An alternative list using the complex exponential, $\exp(\pm im_L\phi)$, has also been used. Some examples of the dependences of the probability density on r and on the azimuthal angle θ are given in Fig.156.2. These spatial dependences are particularly helpful for understanding the electronic energy band properties in semiconductors and solids described in in the latter sections (16n, 17n, 18n) of this chapter. They are also the basis for the chemical bond model of molecules and crystals.

We shall summarize the key features of the electron wavefunctions in hydrogen atom given in the Table 156.1 and of their radial and angular dependences given in Figs.156.2(a)-(c). The electron wavefunctions in hydrogen atom are frequently called **hydrogen wavefunction**. This terminology is somewhat misleading since it is not the wavefunction of a hydrogen atom. It is the wavefunction of an electron bound by the Coulomb force from the proton in the hydrogen atom. We shall however frequently employ this customary usage, viz., hydrogen wavefunction. Some of the key properties are itemized in the paragraphs after the following table and figures.

TABLE 156.1
Normalized Electron Wavefunctions in Hydrogen Atom

n	L	m	
1	0	0	$\psi_{1s} = \frac{1}{\sqrt{\pi}}\left(\frac{Z}{a_0}\right)^{3/2} e^{-\rho}$
2	0	0	$\psi_{2s} = \frac{1}{4\sqrt{2\pi}}\left(\frac{Z}{a_0}\right)^{3/2} (2-\rho)e^{-\frac{\rho}{2}}$
2	1	0	$\psi_{2p_z} = \frac{1}{4\sqrt{2\pi}}\left(\frac{Z}{a_0}\right)^{3/2} \rho e^{-\frac{\rho}{2}} \cos\theta$
2	1	±1	$\psi_{2p_x} = \frac{1}{4\sqrt{2\pi}}\left(\frac{Z}{a_0}\right)^{3/2} \rho e^{-\frac{\rho}{2}} \sin\theta\cos\varphi$
			$\psi_{2p_y} = \frac{1}{4\sqrt{2\pi}}\left(\frac{Z}{a_0}\right)^{3/2} \rho e^{-\frac{\rho}{2}} \sin\theta\sin\varphi$
3	0	0	$\psi_{3s} = \frac{2}{81\sqrt{3\pi}}\left(\frac{Z}{a_0}\right)^{3/2} (27-18\rho+2\rho^2)e^{-\frac{\rho}{3}}$
3	1	0	$\psi_{3p_z} = \frac{2}{81\sqrt{\pi}}\left(\frac{Z}{a_0}\right)^{3/2} (6\rho-\rho^2)e^{-\frac{\rho}{3}} \cos\theta$
3	1	±1	$\psi_{3p_x} = \frac{2}{81\sqrt{\pi}}\left(\frac{Z}{a_0}\right)^{3/2} (6\rho-\rho^2)e^{-\frac{\rho}{3}} \sin\theta\cos\varphi$
			$\psi_{3p_y} = \frac{2}{81\sqrt{\pi}}\left(\frac{Z}{a_0}\right)^{3/2} (6\rho-\rho^2)e^{-\frac{\rho}{3}} \sin\theta\sin\varphi$
3	2	0	$\psi_{3d_{z^2}} = \frac{1}{81\sqrt{6\pi}}\left(\frac{Z}{a_0}\right)^{3/2} \rho^2 e^{-\frac{\rho}{3}} (3\cos^2\theta - 1)$
3	2	±1	$\psi_{3d_{xz}} = \frac{\sqrt{2}}{81\sqrt{\pi}}\left(\frac{Z}{a_0}\right)^{3/2} \rho^2 e^{-\frac{\rho}{3}} \sin\theta\cos\theta\cos\varphi$
			$\psi_{3d_{yz}} = \frac{\sqrt{2}}{81\sqrt{\pi}}\left(\frac{Z}{a_0}\right)^{3/2} \rho^2 e^{-\frac{\rho}{3}} \sin\theta\cos\theta\sin\varphi$
3	2	±2	$\psi_{3d_{x^2-y^2}} = \frac{1}{81\sqrt{2\pi}}\left(\frac{Z}{a_0}\right)^{3/2} \rho^2 e^{-\frac{\rho}{3}} \sin^2\theta\cos 2\varphi$
			$\psi_{3d_{xy}} = \frac{1}{81\sqrt{2\pi}}\left(\frac{Z}{a_0}\right)^{3/2} \rho^2 e^{-\frac{\rho}{3}} \sin^2\theta\sin 2\varphi$

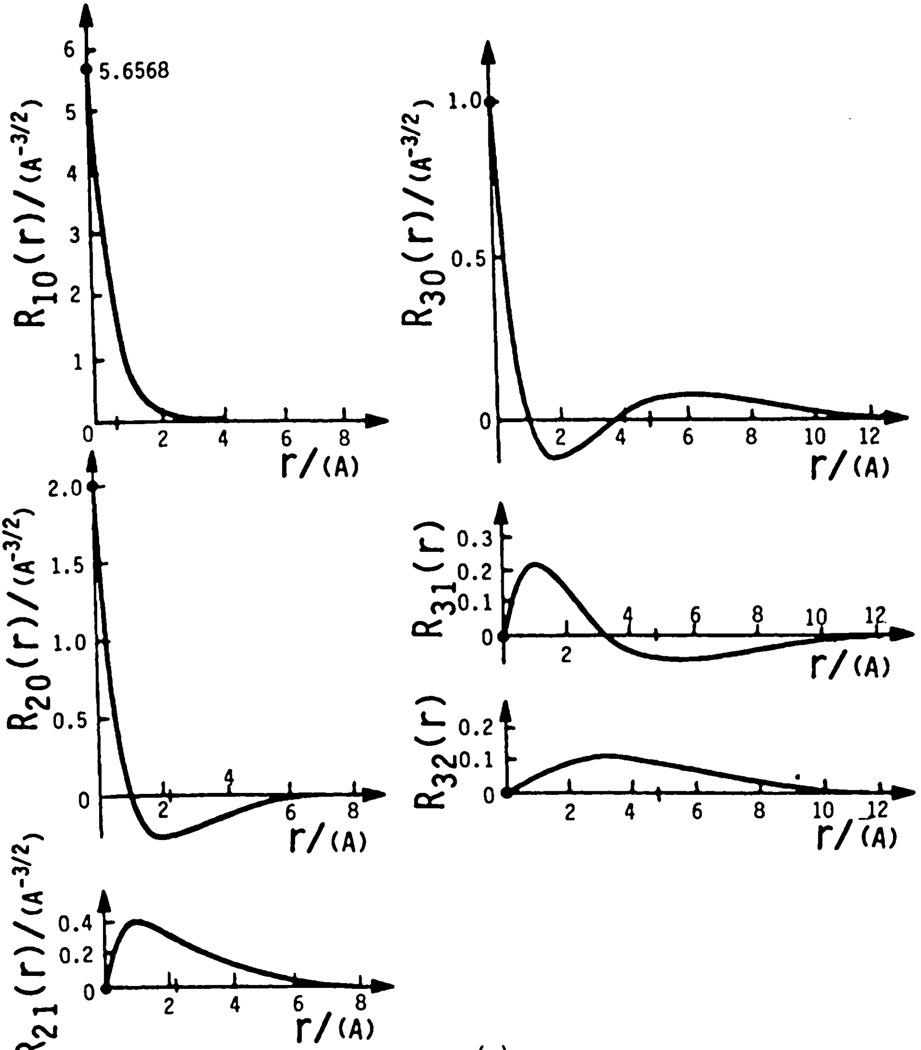
Legend:

Normalized distance $\rho = + Zr/a_0$

Potential Energy $V(r) = - Zq^2/4\pi\varepsilon_0 r$ (Point charge of +Zq)

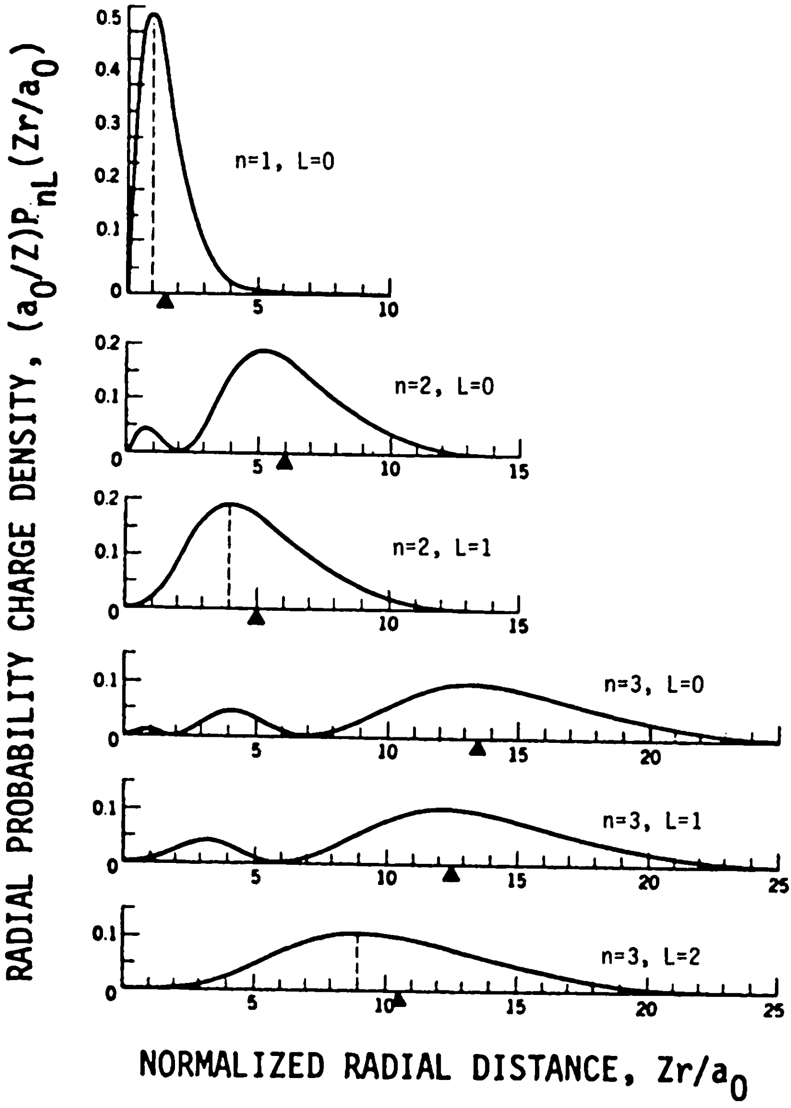
Bohr radius $a_0 = + 4\pi\varepsilon_0\hbar^2/mq^2 = + 0.5292$ Angstrom

Bound state energy $E_n = + Z^2q^2m/[2(4\pi\varepsilon_0)^2\hbar^2n^2] = - 13.605(Z/n)^2$ (eV)



(a)

Fig.156.2 The spatial variations of the electron wavefunctions in hydrogen atom whose nucleus or proton is located at the origin of the coordinate. (a) Radial wavefunction. (b) Radial charge density. (c) Polar graphs of the directional dependence of the probability density, $|\Theta_{Lm}(\Theta)|^2$.



(b)

Fig.156.2 The spatial variations of the electron wavefunctions in hydrogen atom whose nucleus or proton is located at the origin of the coordinate. (a) Radial wavefunction. (b) Radial charge density. (c) Polar graphs of the directional dependence of the probability density, $|\Theta_{Lm}(\Theta)|^2$.

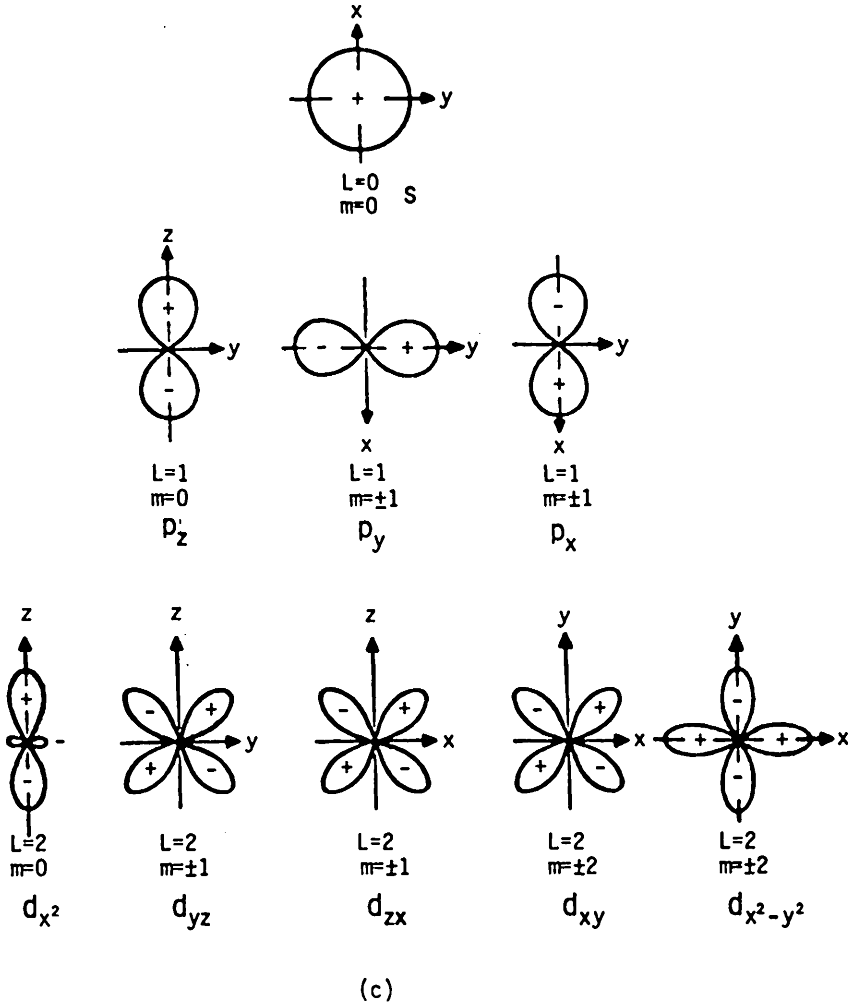

x
y
+
L=0
m=0
s
z
y
+
-
L=1
m=0
p$_z$
x
y
-
+
L=1
m=±1
p$_y$
x
y
-
+
L=1
m=±1
p$_x$
z
+
-
-
L=2
m=0
d$_{x^2}$
z
y
L=2
m=±1
d$_{yz}$
z
x
L=2
m=±1
d$_{zx}$
y
x
L=2
m=±2
d$_{xy}$
y
x
L=2
m=±2
d$_{x^2-y^2}$
(c)

Fig.156.2 The spatial variations of the electron wavefunctions in hydrogen atom whose nucleus or proton is located at the origin of the coordinate. (a) Radial wavefunction. (b) Radial charge density. (c) Polar graphs of the directional dependence of the probability density, $|\Theta_{Lm}(\theta)|^2$.

(A) Label the Wavefunction

The wavefunction is labeled or indexed by the three quantum numbers, n, L, and m_L, as subscripts: (Some authors use m, such as Slater.)

$\psi_{nLm}(r,\theta,\phi)$.

(B) Separate the Wavefunction into Three Components.

The wavefunction is written as a product of three functions: $R_{nL}(r)$, $\Theta_{Lm}(\theta)$ and $\Phi_m(\phi)$ and given by

$$\psi_{nLm}(r,\theta,\phi) = R_{nL}(r)\cdot\Theta_{Lm}(\theta)\cdot\Phi_m(\phi).$$

(C) Tabulate the Radial and Angular Wavefunctions for Z Nuclear Charge

The lower energy wavefunctions of different combinations of the three quantum numbers are given in the Table 156.1. Here Z is the magnitude of the positive nuclear charge in the unit of electron charge, q. The Coulomb potential energy of the electron due to the Z units of positive nuclear charge is given by $V(r) = -Zq^2/4\pi\varepsilon_0 r$. The Bohr radius is usually represented by the symbol, $a_0 = 4\pi\varepsilon_0\hbar^2/mq^2$ for the ground state $n=1$ which has the lowest energy, and $a_n = a_0/n^2$ for the excited states at higher energies. In the elementary semiclassical description of the hydrogen atom given by Bohr in section 141, we have used the symbol r_n for the Bohr radii where $n=1,2,3,\ldots$ and r_1 is identical to a_0 used here and in the future.

(D) Illustrate the Ground State or Lowest Energy Wavefunction, 1s

$$\psi_{100}(r,\theta,\phi) = (1/\sqrt{\pi})(Z/a_0)^{3/2}\exp(-Zr/a_0)$$

is the wavefunction of the electron in the state $n=1$, $L=0$, and $m_L=0$. It is known as the ground state wavefunction. It is independent of the angles, θ and ϕ, and only depends on r. Thus, it is **spherically symmetric**. It is also known as the **1s wavefunction** where 1 comes from $n=1$ and s comes from $L=0$.

(E) Illustrate the 3s Wavefunction

As a second illustration, ψ_{300} is the **3s wavefunction**. It is again independent of the angles. This spherical symmetry or independence of angle is a common character of all s-wavefunctions. The difference between 1s and 3s wavefunctions is that there is 1 zero in the 1s and 3 zeros in 3s wavefunctions as illustrated in Fig.156.2(b). Zero at the origin

$r=0$ is counted. The more zeros means more oscillation or more oscillatory and hence higher kinetic energy.

(F) Illustrate the three 3p Wavefunctions

As a third illustration, we pick the ψ_{310}, ψ_{311} and $\psi_{31\text{-}1}$ which are the three 3p wavefunctions written in the spherical coordinate. Their θ dependences are $\cos\theta$ and $\sin\theta$. There is another convention used for the p-type wavefunctions, which is to write them in the 3-dimensional Cartesian coordinate. To get the cartesian forms, one can make a linear combination of the three wavefunctions in the spherical coordinate. The wavefunctions in the Cartesian coordinate is denoted by $3p_x$, $3p_y$ and $3p_z$ which has the form of $3p_x = xF_{31}(r)$, $3p_y = yF_{31}(r)$ and $3p_z = zF_{31}(r)$. Here, $F_{31}(r) = R_{31}/r$.

(G) Relate the four n=3 (3s, 3p$_x$, 3p$_y$, 3p$_z$) Wavefunctions to Electrons in Silicon Semiconductor and in Si Transistors

The notations of the $n=3$, and $L=0$ and $L=1$ wavefunctions, given by 3s and $3p_x$, $3p_y$ and $3p_z$, are particularly convenient to use when we discuss the origin of the electron energy bands in silicon. The reason is that the important energy bands in which the electrons are responsible for the electrical properties of semiconductors and transistors are those formed by the four valence electrons (two 3s and two 3p electrons) in the case of Si. For Ge, these are the 4s and 4p electrons while in diamond, they are the 2s and 2p electrons.

(H) Radial Dependences of the Wavefunction

The radial dependences of the wavefunction, $\psi(r)$, are shown in Fig.156.2(a) for several quantum numbers. The special feature is that the s-type wavefunctions ($L=0$) are all finite or nonzero at the origin while those with angular momentum ($L\neq0$) are zero at the origin. This spatial dependence is rarely described and illustrated graphically in textbooks. However, the spatial dependences are very important in showing the symmetry and magnitude of the wavefunctions, especially if they are or are not zero at the origin. These dependences tell us many important properties of the valence electrons and their effects on the electrical conductivity which we shall illustrate. For example, why is the hole mobility or conductivity always smaller than the electron mobility or conductivity? And why is this accentuated in compound semiconductors and increasingly so in oxide semiconductors and insulators?

(I) Radial Dependences of the Probability Density Function

The radial dependences of the wavefunction, R(r) in Fig.156.2(a), are rarely given, but the probability density function $P(r)=r^2R^2(r)$ are almost always given in textbooks. Some samples for the lowest energies are shown in Fig.156.2(b). The radial distance in the figure is normalized to the Bohr radius $a_0=0.53A$ and the probability is normalized to (Z/a_0).

The radial probability density function is the fundamental basis of the concept of chemical bonds (covalent bonds, ionic bonds, hydrogen bonds, etc.) which has been used by chemists since Linus Pauling began to systematically classify the chemical properties of materials in the 1930's. The outermost and also the largest peak of the radial probability density function is designated as the **unpaired** or **dangling bond length.** The total bond length of two electrons (one from each of the two adjacent atoms) forming an electron-pair bond is the sum of the two unpaired bond lengths. It gives a measure of the inter-atomic distance of two atoms in a diatomic molecule. The radial probability density function is also the basis used by modern solid state physicists and chemists as well as device engineers to describe material and device properties based on the covalent bonds, the impurity and defect bonds, as well as the ruptured or dangling unpaired bonds. This dangling bond determines the performance and reliability-durability of transistors. Thus, a precise definition of the probability functions is desirable which is given as follows.

The **total probability** of finding an electron in a volume element d^3r is $\psi^*\psi d^3r$. In spherical coordinate, $d^3r=r^2dr\cdot\sin\theta d\theta\cdot d\phi$. So the total probability of finding the electron in the volume element d^3r is

$$P(r,\theta,\phi)d^3r = [R_{nL}{}^2(r)r^2dr]\cdot[\Theta^2\sin\theta d\theta][\Phi^2d\phi]$$

where $P(r,\theta,\phi)$ the **volume probability density** (probability per unit volume). The first bracketed term is defined as the **radial probability** which can be written as

$$P_{nL}(r)dr = R_{nL}{}^2(r)r^2dr.$$

Thus, the **radial probability density** is given by

$$P_{nL}(r) = r^2R_{nL}{}^2(r). \qquad (156.7)$$

It is the **probability per unit length** in the radial direction dr. In Fig.156.2(b), the dimensionless or normalized radial probability density, $(a_0/Z)P_{nL}(r)$ is plotted. Note that the largest maximum occurs at larger r for larger n and L but for a given n, the position of the largest and also **outermost maximum** decreases with increasing L (see the lowest three curves).

(J) Angular Dependence of the Probability Density Function

The two angular components, θ and Φ, are combined since Φ's amplitude is independent of the polar angle ϕ: $\Phi^*\Phi = \exp(-im_L\phi)\exp(im_L\phi) = 1$. The angular dependence all comes from the azimuthal angle θ. The angular probability density is defined as the **probability per unit solid angle**. The differential solid angle is given by

$$d\Omega = \sin\theta d\theta d\phi \qquad (156.8)$$

and the angular probability is given by

$$P_{Lm}(\theta)d\Omega = P_{Lm}(\theta,\phi)\sin\theta d\theta d\phi = [\theta^*\theta\sin\theta d\theta][\Phi^*\Phi d\phi] = \theta^*\theta d\Omega.$$

Thus, the angular probability density is given by

$$P_{Lm}(\theta) = \theta_{Lm}^*(\theta)\cdot\theta_{Lm}(\theta).$$

It is dimensionless and normalized. It is a function of θ and it also depends on the quantum numbers L and m_L but not on the principal quantum number n.

Drawings in Fig.156.2(c) give the dependence of the angular probability density plotted in the cross-sectional plane such as $z=0$ or x-y plane, and in polar coordinate (ρ,θ) where $\rho=\sqrt{(x^2+y^2)}$. These figures are specially selected from the literature since some published figures are in error and others use different definitions and normalization constants.

The case of $L=0$ (s-state) is a circle since there is no angular dependence, i.e. it is spherically symmetric.

For $L=1$ (p-states), there are three linearly independent, differently oriented wavefunctions. **Linearly independent** means that two of these three functions cannot be combined to form the third. These three are shown in Fig.156.2(c), for m_L(or m used by Slater)$=0$, $m_L=+1$, and $m_L=-1$. From Table 156.1 of the wavefunctions, we see that for $L=1$ and $m_L=0$, the angular wavefunction has the form of $\cos\theta$. Thus, the angular probability density is $\cos^2\theta$ which has its maximum amplitude pointing in the $+z$ and $-z$ directions as indicated by the p_z figure in Fig.156.2(c). For $L=1$ and $m_L=\pm1$, $\theta=\sin\theta$ and $P(\theta)=\sin^2\theta$ which has its maximum amplitude pointing in the $+y$ and $-y$ directions as indicated by the p_y figure in Fig.156.2(c). If we make this choice for the second p-function, then the third p-function, p_x, would be $\sin^2\theta$ pointing in the $+x$ and $-x$ directions perpendicular to the paper of Fig.156.2. These are the three p-states: p_x, p_y and p_z. Angular dependences of the five d-orbitals are also shown in Fig.156.2(c). The signs, + and -, inside the lobes indicate whether the wavefunction is positive or negative. The sign is particularly valuable when calculating overlap energy of electrons located on adjacent atoms, giving the notion of **bonding** and **antibonding**. The angular dependences have been presented in polar coordinates such as the four examples given in Fig.156.2(c). These polar plots are particularly useful for studying the chemical bond model of electronic solids such as silicon, silicon dioxide, and other semiconductors, since they show bond angle and overlap of the electron wavefunctions from adjacent atoms. The amount of overlap gives the bond strength.

(K) Degeneracy of Electron Wavefunctions and Electron States (Configuration Degeneracy and Spin Degeneracy)

The solution of a partial differential equation is termed **degenerate** when there are many linearly independent solutions all having the same eigenvalue or allowed energy. In the hydrogen atom problem, the allowed energy is designated by the principal quantum number, n. We have seen from Table 156.1 that for a given quantum number n, we may have several linearly independent wavefunctions of different spatial variations. For example, for the state $n=2$ we have a group of four degenerate wavefunctions or solutions, 2s, $2p_x$, $2p_y$ and $2p_z$; and for $n=3$, we have a group of nine degenerate wavefunctions, 3s, $3p_x$, $3p_y$, $3p_z$, $3d_{z^2}$, $3d_{yz}$, $3d_{zx}$, $3d_{xy}$, and $3d_{x^2-y^2}$. Linear independence means that one cannot form one of the wavefunctions by a linear combination of the remaining wavefunctions in the group. In these two examples, the multiplicity of linearly independent wavefunctions, all with the same energy but all with different spatial variations, is known as **degeneracy** or more precisely, **quantum configuration degeneracy** since they have different spatial configuration but the same energy.

The adjective 'degenerate' has an entirely different meaning when it is used in equilibrium statistical mechanics to analyze the properties of many interacting particles in a solid. When the electron density is high, such as in a high-conductivity semiconductor or metal, the statistical distribution of the kinetic energy of the electrons is known as **degenerate** and it is represented by the **Fermi** (or **Fermi-Dirac**) function. When the electron density is low, less than about 10^{18}cm^{-3} in Si at room temperature, the statistical distribution of the kinetic energy of the electrons is known as **nondegenerate** and it is given by the **Boltzmann** (or exponential) function. To distinguish the quantum and statistical mechanical usages clearly, a qualifier may be used: **carrier concentration degeneracy**, **concentration degeneracy**, or **statistical distribution degeneracy** for statistical mechanics; and **configuration degeneracy**, or **quantum configuration degeneracy** for quantum mechanics.

In addition to the quantum configuration degeneracy just described, there is a second degeneracy from quantum mechanics. This comes from the electron spin because the electron may have two spin orientations (up or down) at the same energy. Thus, the **spin-degeneracy** increases the total quantum degeneracy by a factor of 2 over the configuration degeneracy. For example, for the n=2 principal quantum number at an energy E_2, there are 8 degenerate states (4-fold configuration degeneracy each with a 2-fold spin-degeneracy, giving 4x2=8). Sometimes, we say that n=2 in hydrogen atom is an 8-fold degenerate energy level. The spin degeneracy can be lifted or removed by spin-orbit interaction and relativistic effects, so that the states with the same n but different L may have slightly different energies. Then, including the two spin states, **a distinct electron quantum state is specified by four quantum numbers**, n, L, m_L, and m_s where $m_s=\pm 1/2$ is the spin quantum number.

(L) Spectroscopic Notations

To simplify the discussion and notation of spectral lines of light emitted by an excited hydrogen atom (and also heavier elements), spectroscopists have adopted a letter notation to indicate the many-fold degenerate states. They have designated the principle (n) and orbital (L) quantum numbers by numbers designated the orbital quantum number. The convention is K-shell for n=1; L-shell for n=2; M-shell for n=3; etc.; and s for L=0; p for L=1, d for L=2, f for L=3, etc. The notation $(nL)^g$ is also used to indicate that there are g-number of electrons occupying a particular group of quantum states whose quantum numbers are n and L. We have so far talked only about the one-electron atom, the hydrogen atom, hence g=1. In the following two sections, we will describe the adaptation of this one-electron spectroscopic notation and wavefunction results to the many-electron heavier atoms and to solids which contain many atoms and many more electrons.

For the hydrogen atom, if the one electron is in the n=0 or 1s state, we write this as $(1s)^1$ and this is also known as the ground state. If this electron is in the n=2 and L=0 state, we represent it by $(2s)^1$ and the hydrogen atom is now in an excited state. In another excited state such as n=3, L=1, it is designated by $(3p)^1$. The photon emitted by the electron making a quantum mechanical transition from the (3p) state to the ground state (1s) can be written as $(3p)^1 \rightarrow (1s)^0$ which is the second line observed on the photographic recording of the Lyman series of the spectral lines emitted by a hydrogen atom.

An especially bothersome point for beginners and sometimes veterans concerns the fundamental meaning of the many allowed energy levels and wavefunctions just described for one electron bound to a proton nucleus.

The key point is that these are allowed solutions for **one electron**. Once an electron occupies one of the allowed solutions or energy levels, all the remaining energy levels and solutions no longer exist. For a second electron, a new and completely different set of allowed energy levels and wavefunctions will appear if a set of bound solutions exist for binding the second electron. The new set is different from the solution in a **neutral hydrogen** (one electron around one proton) since the new solution is the solution of two electrons bound to one proton, i.e., it is the solution of a **negatively charged hydrogen ion**, H^- or $H:^-$. The energy required to release one of the two bound electrons from H^- is the **electron affinity**, depicted by the **transition equation**, $H:^- \rightarrow H\cdot + e^-$. For hydrogen, this energy is 0.75415eV in theory and 0.756eV measured experimentally. These values are very different from the ionization, 13.6eV, of a hydrogen atom, $H\cdot \rightarrow H^+ + e^-$. For Si atom, the electron affinity is 1.39eV; for Si crystal, it is 4.02eV; and for SiO_2 solid, it is 0.9eV.

160 ELECTRON CONFIGURATIONS IN MANY-ELECTRON ATOMS

The one-electron energy model described in section 156 for the hydrogen atom is now extended to atoms with many electrons. We shall see in the next sections, 17n, that the one-electron model is a very powerful and vivid model to simplify the concepts and physics of electronic conduction in semiconductors and transistors that contain very many electrons and very many protons or atoms. Thus, we shall first extend the one-electron hydrogen model to many-electron atoms. In section 161, the simplest many-electron atoms, one proton with two electrons or the negatively charged hydrogen atom, is described in order to introduce the important fundamental concepts and basic applications of the energy level diagrams level. In particular, we shall show how the one-electron energy diagram can be developed for a two-electron atom and how it is used to give precise values of the energies during an electronic transition such as the emission of light described in section 141 for the one-electron Bohr atom. In section 162, the one-electron energy model for atoms with more than one proton and many electrons is then described.

161 The Negatively Charged Two-Electron Hydrogen Atom

The quantum mechanical model of the hydrogen atom was described in section 156. It has one proton and one electron, and is a two-particle, two-body or one-electron problem. The bound solutions of the Schrödinger equation (the localized wavefunctions and discrete energy levels) are those of the single electron bound to the Coulomb potential well of the positively charged hydrogen nucleus, the proton. The solutions are labeled by four quantum numbers, n, L, m_L, and s. The square of the wavefunction is the position probability of finding the electron around the proton at a particular allowed total energy, E_n. They are the bound electron solutions in a neutral hydrogen.

If a second electron is trapped to the proton, there are then two electrons and a proton. It is then a three-particle, three-body or two-electron atom. It is a negatively charged hydrogen atom or ion. The second electron is bound to a rather weak attractive potential well from the neutral hydrogen atom. The weak attractive or positive potential seen by the second electron is the residual due to incomplete spatial cancellation of the positive proton charge by the negatively charged orbiting electron. Thus, the wavefunctions and energy levels of the second bound electron are very different from that of the one-electron solutions around a proton. The one-electron solutions in fact no longer exist: they are modified by the electron-electron repulsion between the two electrons. Thus, the total ground state energy of the two electrons, E_{20} (E_{jn} where j=number of electron and n=principal quantum number), is larger (or less negative) than that of one electron, E_{10}, or $E_{20} > E_{10} = -13.6\text{eV}$. The difference, $\chi = E_{20} - E_{10}$, is known as the **electron affinity of a neutral atom.** For hydrogen, it has been determined experimentally (0.756eV) and theoretically (0.75415eV), giving $E_{20} = E_{10} + \chi = -13.605 + 0.75415 = -12.850\text{eV}$. Thus, the two electrons are sitting at an energy level of $-12.850/2 = -6.425\text{eV}$ with opposite spin.

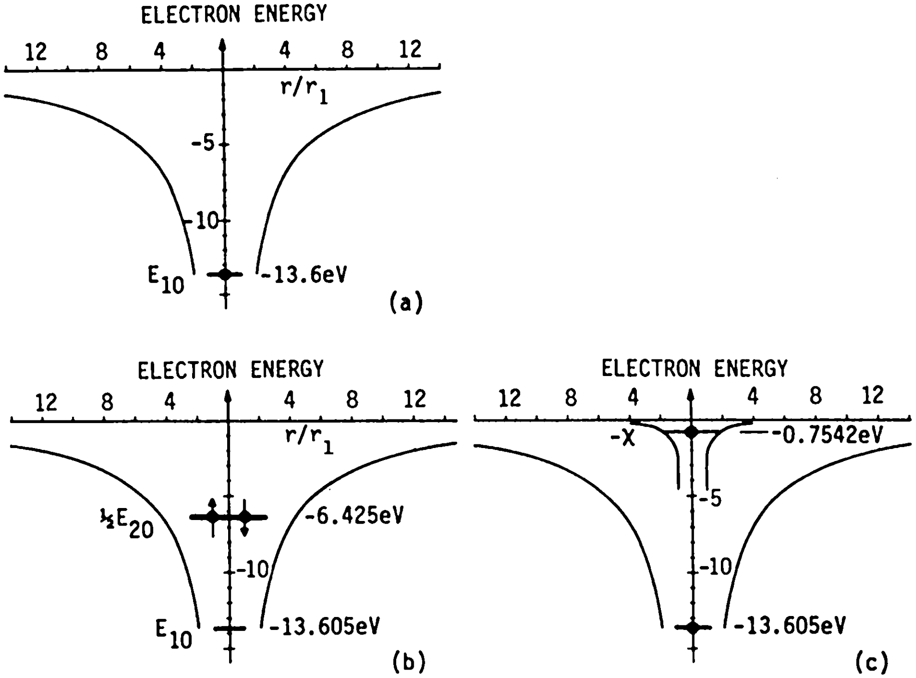


Fig.161.1 The electron energy diagram of neutral and negatively charged hydrogen atoms. **(a)** Energy level diagram of a neutral hydrogen with one electron trapped to a proton. **(b)** Energy level diagram of a negatively charged hydrogen with two trapped electrons. **(c)** Energy change or transition energy diagram of a negatively charged hydrogen with two trapped electrons.

The electron energy diagram for the neutral hydrogen atom is shown in Fig.161.1(a) where the electron potential energy in two continuous curved lines is the exact Coulomb energy from the proton, $-q^2/(4\pi\varepsilon_0)$.

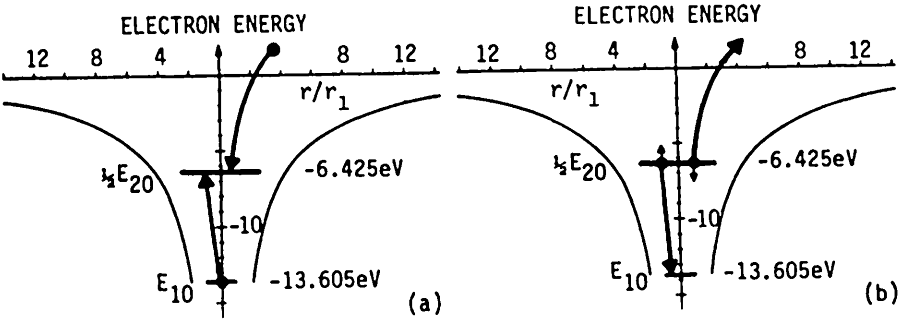

ELECTRON ENERGY
12
8
4
8
12
r/r₁
½E₂₀
-6.425eV
-10
E₁₀
-13.605eV
(a)
ELECTRON ENERGY
12
8
4
8
12
r/r₁
½E₂₀
-6.425eV
-10
E₁₀
-13.605eV
(b)

Fig.161.2 Electronic affinity transitions in a neutral and negatively charged hydrogen atom illustrated by the energy level diagram. (a) Electron capture by a neutral hydrogen atom. (b) Electron emission by a negatively charged hydrogen atom.

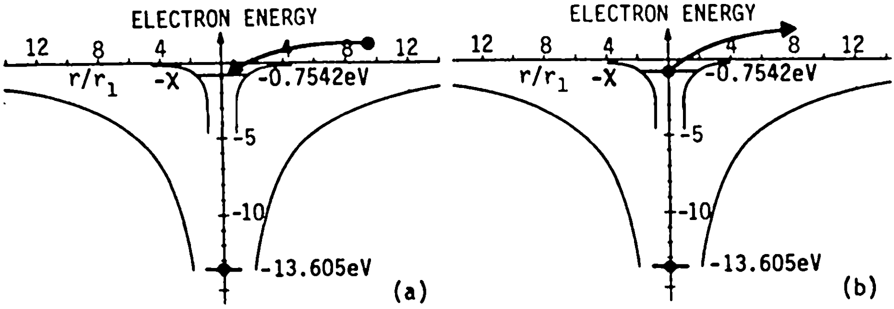

ELECTRON ENERGY
12
8
4
8
12
r/r₁
-X
-0.7542eV
-5
-10
-13.605eV
(a)
ELECTRON ENERGY
12
8
4
4
8
12
r/r₁
-X
-0.7542eV
-5
-10
-13.605eV
(b)

Fig.161.3 Electronic affinity transitions in a neutral and negatively charged hydrogen atom illustrated by the transition energy diagram or energy change diagram. (a) Electron capture by a neutral hydrogen atom. (c) Electron emission by a negatively charged hydrogen atom.

For the negatively charged hydrogen atom containing two electrons, one can draw two energy diagrams. Figure 161.1(b) shows the two-electron energy diagram where the potential energy curve is the exact Coulomb electron potential energy from the proton, $-q^2/(4\pi\varepsilon_0 r)$, and the energy level is the exact bound state energy of the ground state solution of the two-electron one-proton Schrödinger equation. Figure 161.1(c) shows the one-electron energy diagram where the smaller and squarish potential energy curve is the effective potential energy of the second electron (the dot at -0.75415eV), known as the one-electron potential energy. The one-electron potential energy can be only conceptually sketched and hence represented by broken lines in Fig.161.1(c). It is only conceptual or pictorial because the potential energy seen by the second electron is a nonlocal potential owing to the orbital motion of the first electron that screens the proton charge (Bohr model). This orbital motion results in a space-time dependent effective potential (stationary proton minus orbiting first electron) seen by the second electron. The one-electron energies or energy 'levels' in the one-electron energy diagram are the electron transition energies or electron energy changes during an electron transition between two states. They are not the eigenvalues or allowed value of the energy in the Schrödinger equation but are computed from the E_n's. Thus, the one-electron energy diagram should be termed the transition energy diagram or energy change diagram, however, it is seldom recognized as such. (Energy change in contrast to energy level and energy band described in sections 17n and 18n.) The conventional usage, energy level diagram, tends to obscure the fundamental physics and bases underlying the one-electron model and gives the erroneous notion that the energies labeled are the energy eigenvalues of the Schrödinger equation.

The sketched one-electron energy diagram is imprecise. However, it is graphically more vivid than the two-electron energy diagram to demonstrate electronic transitions, such as the emission and absorption of light, since the transition energies are accurately and explicitly labeled. Its vividness is demonstrated by comparing Figs.161.2(a) and (b) using the two-electron energy diagram, with Figs.161.3(a) and (b) using the one-electron energy diagram. Figures 161.2(a) and 161.3(a) show the trapping transition of the second electron being captured by a neutral hydrogen atom. Figures 161.2(b) and 161.3(b) show the detrapping transition of the second electron being emitted or released by the negatively charged two-electron hydrogen atom. The cooperating photon for the two electronic transition is shown as a sinusoidal arrow. It is evident that the one-electron energy diagrams are simpler to visualize because the relevant transition energies are shown on the diagrams. However, one must remember that the one-electron energy diagram (energy change diagram or transition energy diagram) does not give the bound state energy levels or eigenvalues obtained from the two-electron one-proton Schrödinger equation. Instead, it gives the transition energies, for example, 0.75415eV is the energy required to remove the second electron (electron affinity) and 13.602eV is the energy required to remove the first electron after the second electron had been removed (ionization potential).

This exposition can be continued to a double negatively charged hydrogen ion which is a 3-electron and 1-proton or 4-particle problem. However, there is no assurance that a bound solution of the Schrödinger equation exists.

162 Many-Proton and Many-Electron Atoms

For a heavier atom with many protons in the nucleus and many electrons bound to the attractive potential well by the electrostatic or Coulomb force, there is no simple analytical solution except the case of a completely ionized heavier atom which has no electrons bound to its nucleus. In this case, the solutions of the one-electron one-proton hydrogen atom given in section 156 and Table 156.1 will apply if Z is equated to the charge of the bare nucleus (bare means no electrons around it) or the number of protons in the nucleus. The basis is that the nucleus is so much smaller than the Bohr orbits that even when it contains many protons and neutrons, it can still be considered as a **point charge**. Thus the solutions of the hydrogen atom, which is a **point charge model**, apply truthfully to a **bare** heavier nucleus. Corrections are needed for relativistic effects, such as the relativistic mass and the spin-orbit force, since electron in the heavier atom is moving much faster in a much smaller orbit than in hydrogen.

The notations of the one-electron one-proton model of the neutral hydrogen described in section 156 are found to be very useful as a baseline to describe the electron states in a **neutral many-electron** many-proton atom. The numerical values of the one-electron wavefunctions and energies in the one-proton hydrogen atom are no longer correct in a many-electron many-proton atom. They are different because the nucleus of a heavier atom is surrounded by many electrons which shield the Zq nuclear charge **incompletely**. The shielding is incomplete because the surrounding electron orbits around the nucleus, hence, it does not cancel exactly one positive nuclear charge. This was demonstrated in the two-electron one-proton case in the preceding section. But the concept of enumeration of the electron states just developed for a bare nucleus or a point charge (the neutral hydrogen atom) is still valid. Thus, in describing a neutral many-electron atom, we will speak of the many electron states and shells or orbits using the same notations as that used in the one-electron atom, such as 1s, 2s, 2p, 3s, 3p, 3d, etc. They provide an informative label of the electron states and they suggest the resemblance of the many-electron wavefunctions to the one-electron wavefunction.

In addition to the inapplicability of numerical values of the one-electron wavefunctions and energies to the many-electron atoms, there is a major difference owing to the presence of more than one electron. In a many-electron atom, many of the states or energy levels can be occupied by the electrons present. The occupation of a state is governed by Pauli's exclusion principle which states that a quantum state (include spin state) can be occupied by no more than one electron. Thus, the notation we described for the occupation of a group of quantum states by a group of electrons in paragraph (K) of section 156, $(nL)^g$, is now more colorful and quite useful for a many-electron atom. The maximum number of electrons that can occupy a group of quantum states or the maximum value of the index g must now be limited by Pauli's exclusion principle. A count of the numbers in Table 156.1 will show that for each L there are 2L+1 allowed-values of m_L. This was also stated as a result of the angular dependent part of the Schrödinger equation, (154.5). For each of these magnetic quantum states, we have two spin states. Thus, the total number of quantum states for a given L is 2(2L+1). Pauli's principle restricts one electron per state. Thus, the maximum value of g is, $g_{max}=2(2L+1)$ or $g \leq 2(2L+1)$ and this group of quantum states is g_{max}-fold degenerate. The following example gives an illustration.

$(1s)^2(2s)^1(2p)^3$ has the following meaning:

$(1s)^2$	means that the two 1s states are occupied by two electrons.
$(2s)^1$	means that one of the two 2s states is occupied by an electron while the other is unoccupied or empty.
$(2p)^3$	means that three of the six 2p states are each occupied by one electron and three of the six 2p states are not occupied or unoccupied.

In the above configuration, the atom is in an **excited state** since one of the two 2s states is not occupied while three of the six higher energy 2p states are occupied. The ground state configuration is $(1s)^2(2s)^2(2p)^2$. It is the ground state of the neutral carbon atom.

As a second example, the ground state configuration or **ground state electronic configuration** of a neutral silicon atom is $(1s)^2(2s)^2(2p)^6(3s)^2(3p)^2$. The ground state configuration of a singly ionized silicon atom or singly charged silicon ion is $(1s)^2(2s)^2(2p)^6(3s)^2(3p)^1$ in which one of the two 3p electrons has been released or removed from a neutral silicon atom. The configuration of $(1s)^2(2s)^2(2p)^6(3s)^2(4s)^1$ is that of a Si+ ion in an excited state which can decay to the ground state by emitting light.

Wavefunctions and the radii of the maximum radial probability (charge) density have been computed numerically for all the elements on high speed number-crunching computers by self-consistent numerical iteration procedures. The central problem is that there are many electrons whose electron-electron repulsive potential energy is no longer the simple spherically symmetric 1/r-type from a positive point charge of the nucleus. The one-electron energies, the wavefunctions, and the positions of the maximum radial probability density of the **neutral** many-electron atoms have been computed by Herman and Skillman and by Waber and Cromer and the results are tabulated. The energies and the locations of the maxima of radial charge densities are given in two tables, Tables 162.1 and 162.2, on the following two pages. The radii can be used to produce a sketch of the charge cloud such as the picture for Si in Fig.171.1 given in the next section.

Although the results of the one-electron hydrogen energies given by the Bohr atom model in section 141 and the Schrödinger equation solution in section 156 do not give the exact electron energies in many-electron atoms nor predict them, as revealed in Table 162.1, the general trend of how the one-electron-energy changes with elements can be understood qualitatively. The key is the concept of an **effective nuclear charge**, Z(r), in the potential energy of the j-th electron, $V(r) = -q^2Z(r)/4\pi\varepsilon_0 r$. This effective nuclear charge qualitatively and semi-quantitatively takes into account the **shielding** of the nuclear charge, +Zq, by the other j-1 electrons in a j-electron atom.

To illustrate the shielding of the nuclear charge by the core, we describe several numerical examples each contains a sequence of atoms. We use the one-electron energy and radius tables given by Tables 162.1 and 162.2.

Consider the sequence of atoms B, C, N, 0, P, F, Ne whose one-electron energies are given in Table 162.1. The one-electron energy for one of the 2p electrons increases from -0.2449, -0.33015, -0.42225, -0.52045, -0.6251 to -0.7335 in Hartree energy unit which is twice the Rydberg-unit or 1 Hartree = 2x13.6eV = 27.2eV. This decreasing trend of one-electron energy is expected. Since as the nuclear charge increases from Z=5 for Boron to Z=1O for Ne with the corresponding addition of 2p electrons from $(2p)^1$ for Boron to $(2p)^6$ for Ne, we would expect the added electrons to the 2p level to be less effective in shielding out the added nuclear charge. This decreasing shielding effectiveness is because the 2p electron is so far out from the nucleus. Thus, the effective nuclear charge, Z(r), will increase instead of being constant when the true nuclear charge increases from Z=5 to Z=10. Larger effective nuclear charge increases the attractive force and gives stronger binding of the added electron to the nucleus. Thus, the binding energy of the 2p electron decreases to larger negative values (or stronger binding) as the nuclear charge is increased from Boron (Z=5) to Neon (Z=10).

The above concept of **incomplete shielding** of the nuclear charge by the **inner shell** or **core electrons** can also be demonstrated by another sequence of atoms such as the ones with one outer shell s-electron: H(1s), Li(2s), Na(3s) and K(4s). The corresponding one-electron energies for the single outer-shell electron from Table 160.1 are: -0.500(H), -0.20195(Li), -0.18885(Na), and -0.1543(K). The predicted values using the simple, complete shielding model of one electron around a point charge of Zq is given by $E_n = -0.500/n^2$ (Hartree unit). Thus, the complete shielding model would give $E_1 = -0.500$, $E_2 = -0.125$(Li), $E_3 = -0.05556$(Na), and $E_4 = -0.03125$(K). These complete shielding results and the exact results from Table 162.1 are tabulated Table 162.3 for comparison.

TABLE 162.1 ONE-ELECTRON ENERGIES IN NEUTRAL ATOMS [Negative values using $V(r=\infty)=0$ as reference and configurations of the out-shell electrons listed. Energies calculated by Herman and Skillman, and tabulated in Hartree unit (1 Hartree = 2 Rydberg = 27.210 eV) by J. C. Slater in **Quantum Theory of Matter**, 2nd ed. (1968), pp.145 and 150, McGraw-Hill.]

		1s	2s	2p	3s	3p	3d	4s	4p
H	$1s$	0.5000							
He	$1s^2$	0.8605							
Li	$2s$	2.199	0.20195						
Be	$2s^2$	4.349	0.3006						
B	$2s^22p$	7.1865	0.46195	0.2449					
C	$2s^22p^2$	10.689	0.64475	0.33015					
N	$2s^22p^3$	14.8685	0.84795	0.42225					
O	$2s^22p^4$	19.728	1.0720	0.52045					
F	$2s^22p^5$	25.269	1.31745	0.6251					
Ne	$2s^22p^4$	31.495	1.584	0.7355					
Na	$3s$	39.025	2.3615	1.3345					
Mg	$3s$	47.475	3.276	2.072	0.25255				
Al	$3s^23p$	56.83	4.3575	2.9735	0.36225	0.1791			
Si	$3s^23p^2$	67.02	5.5435	3.977	0.49875	0.2401			
P	$3s^23p^3$	78.055	6.842	5.090	0.6294	0.30695			
S	$3s^23p^4$	89.945	8.255	6.314	0.7650	0.3781			
Cl	$3s^23p^5$	102.68	9.785	7.6515	0.9062	0.45335			
Ar	$3s^23p^6$	116.27	11.4325	9.1035	1.0534	0.53265			
K	$4s$	131.045	13.530	11.004	1.47605	0.8664	0.1543		
Ca	$4s^2$	146.76	15.8135	13.090	1.9375	1.24115	0.19935		
Sc	$3d^14s^2$	163.145	17.9865	15.065	2.2154	1.44155	0.2654	0.21545	
Ti	$3d^24s^2$	180.385	20.2605	17.1395	2.49065	1.6380	0.31395	0.2289	
V	$3d^34s^2$	198.475	22.6445	19.323	2.76855	1.852	0.35915	0.24095	
Cr	$3d^54s$	217.225	24.9115	21.3895	2.8557	1.8455	0.23945	0.2156	
Mn	$3d^54s^2$	237.235	27.7545	24.025	3.3402	2.2388	0.44295	0.26265	
Fe	$3d^64s^2$	257.905	30.4785	26.542	3.63455	2.4456	0.48125	0.27255	
Co	$3d^74s^2$	279.445	33.324	29.179	3.93925	2.6600	0.52075	0.28235	
Ni	$3d^84s^2$	301.85	36.285	31.93	4.2500	2.878	0.55755	0.29155	
Cu	$3d^{10}4s$	324.85	39.075	34.51	4.317	2.8545	0.37155	0.25455	
Zn	$4s^2$	349.2	42.56	37.775	4.8965	3.3305	0.6291	0.30925	
Ga	$4s^24p$	374.6	46.32	41.315	5.619	3.946	1.0200	0.41855	0.18095
Ge	$4s^24p^2$	400.9	50.255	45.025	6.381	4.5985	1.44505	0.52855	0.23415
As	$4s^24p^3$	428.15	54.365	48.91	7.1875	5.2935	1.91145	0.6377	0.2913
Se	$4s^24p^4$	456.15	58.65	52.975	8.037	6.029	2.4170	0.74765	0.35075
Br	$4s^24p^5$	485.35	63.12	57.215	8.9285	6.8055	2.96225	0.85925	0.41235
Kr	$4s^24p^6$	515.3	67.765	61.625	9.864	7.6245	3.54875	0.9728	0.47595

TABLE 162.2 RADII OF ONE-ELECTRON RADIAL CHARGE DENSITIES

[Radii equated to the outermost maximum of the 1-electron radial charge densities of electrons in the various shells in neutral atoms. Radii in Angstrom Unit or 10^{-8}cm computed by Herman and Skillman, and Waver and Cromer, and tabulated by Slater as cited in Table 162.1.]

	1s	2s	2p	3s	3p	3d	4s	4p
H	0.53							
He	0.291							
Li	0.186	1.586						
Be	0.138	1.040						
B	0.110	0.769	0.776					
C	0.091	0.620	0.596					
N	0.078	0.521	0.488					
O	0.068	0.450	0.413					
F	0.060	0.396	0.360					
Ne	0.054	0.354	0.318					
Na	0.049	0.32	0.278	1.713				
Mg	0.045	0.29	0.247	1.279				
Al	0.042	0.26	0.221	1.044	1.312			
Si	0.039	0.24	0.201	0.904	1.068			
P	0.036	0.22	0.184	0.803	0.918			
S	0.034	0.20	0.169	0.723	0.808			
Cl	0.032	0.19	0.157	0.660	0.723			
Ar	0.030	0.18	0.146	0.607	0.657			
K	0.029	0.17	0.137	0.55	0.593		2.162	
Ca	0.027	0.15	0.130	0.52	0.539		1.690	
Sc	0.026	0.146	0.124	0.48	0.500	0.539	1.570	
Ti	0.025	0.140	0.117	0.459	0.468	0.489	1.477	
V	0.024	0.134	0.110	0.44	0.439	0.449	1.401	
Cr	0.023	0.128	0.105	0.42	0.416	0.426	1.453	
Mn	0.022	0.122	0.100	0.40	0.392	0.389	1.278	
Fe	0.021	0.118	0.095	0.38	0.373	0.364	1.227	
Co	0.020	0.113	0.091	0.356	0.355	0.343	1.181	
Ni	0.019	0.109	0.088	0.34	0.339	0.324	1.139	
Cu	0.018	0.105	0.084	0.33	0.325	0.311	1.191	
Zn	0.018	0.100	0.081	0.32	0.311	0.292	1.065	
Ga	0.017	0.097	0.078	0.303	0.298	0.275	0.960	1.254
Ge	0.016	0.094	0.074	0.29	0.285	0.260	0.886	1.090
As	0.015	0.090	0.070	0.28	0.274	0.246	0.826	0.99
Se	0.015	0.087	0.068	0.27	0.263	0.234	0.775	0.91
Br	0.015	0.085	0.066	0.26	0.253	0.223	0.730	0.840
Kr	0.015	0.083	0.065	0.251	0.244	0.213	0.691	0.79

TABLE 162.3
The Exact and Complete-Shielding Values of the Ground State Energy of the First Four Uni-Valence Elements

ELEMENT	=	**H**	**Li**	**Na**	**K**	**SHIELDING**
Z	=	1	3	11	19	
n	=	1	2	3	4	
E(exact)	=	-.500	-.20195	-.18885	-.1543	**Incomplete**
$E_n = -0.5/n^2$	=	-.500	-.125	-.05556	-.03125	**Complete**
Difference	=	0	-.077	-.133	-.12	

The exact results of Li, Na, and K show larger negative one-electron energy for the s-electron than the simple complete-shielding theory. This shows that shielding of the nuclear charge by the inner core electron is not completely effective. The larger difference for higher Z reflects the fact that the Z-1 core electrons are spread out more in heavier atoms and hence are less effective in screening out the nuclear charges.

The radii of the maximum radial probability (charge) density given in Table 162.2 provides us an estimate of the length of the electron-pair bond when the electrons in the semiconductors are described by the covalent bond model. This is known as the **bond length**. Note that the radii of the 4 outer shell electrons $(3s)^2(3p)^2$ of Si and $(4s)^2(4p)^2$ of Ge are about 1 Angstrom or 10^{-8} cm. Similarly, the averages in GaP and GaAs are also about 1 A. These are consistent with the nearest neighbor atom spacing in these semiconductors, about 2.4A, measured from x-ray diffraction experiments.

In summary, Table 162.1 gives the theoretical one-electron energies of neutral atoms. These are the energies required to remove one electron from a particular energy level or shell in a many-electron atom. It takes into account of the electrostatic or Coulomb forces from all the electrons in the neutral atom but not the electron spin and relativistic mass. It also gives the ionization energy of the neutral atom, which is the minimum energy required to remove the electron in the outermost shell, which is least tightly bound from the neutral atom, causing the atom to become positively charged. **The one-electron energies are in fact the transition energies** required to ionize the atom by releasing an electron from a particular shell to vacuum. The one-electron model of the many-electron atoms described here are used as the starting point to describe the properties of the many-electron and many-atom solids.

170 ELECTRONIC MODELS OF SEMICONDUCTORS AND SOLIDS

There are two ways to describe the electrons and their electrical properties in a solid. These are the qualitative **valence bond model** and the quantitative **energy band model**. They have been used since the 1930's. We shall call these the Bond Model and the Band Model. They are described qualitatively in the following subsections, 17n. The mathematical derivations of the quantitative energy band model are given in the next several sections, 18n.

171 The Bond Model

The bond model was developed by and is popular among chemists. It has given a very lucid pictorial and illustrative mental picture of the electronic and chemical properties of molecules, small and large clusters and chains of atoms, polymers and noncrystalline solids. Since the invention of the transistor in 1948, it has also been a very effective model for demonstrating the electrical properties of electrons and holes in semiconductors and insulators. Although the valence electron and hole particles in the bond model can be analyzed in the same mathematical way as the gas molecules using the kinetic theory of gases, the resultant kinetic or transport coefficients of the electrons and holes (mobility, diffusivity, and lifetimes) cannot be readily related to the fundamental properties of the semiconductor because there is not a simple mathematical link owing to the qualitative nature of the bond model. Some introductory solid-state electronic textbooks have taken the bond-model route to derive the electrical properties of diodes and transistors. This is intellectually unsatisfactory since (i) they must assume that the kinetic or transport coefficients (mobility, diffusivity, and lifetimes) are empirical parameters determined by experiments, and (ii) they cannot give a quantitative description of how these parameters varies with temperature, with the concentration of impurities that are present in the semiconductor, with concentration of crystal defects, with the magnitude of the electric field, and with other technologically important variables. In spite of the inability of the bond model to give quantitative results, it is so useful to illustrate fundamental ideas and so perceptively helpful to formulate qualitative pictures of basic phenomena that it is widely used as a complement to the band model. For example, **bond strength** is a measure of the forces that bind the atoms together in solids but it also measures and indicates the hardiness or endurance of transistors under electrical stress (see section 110).

In elemental semiconductors, such as C(diamond), Si and Ge, the binding force arises from the **electron-pair bonds**. Each bond contains two electrons with opposite spin orientation. The electron distribution is similar around each atom. Thus, the electron pair bond is also known as the **covalent bond** or **homo-polar bond**, in contrast to **hetero-polar** or **ionic bond** in ionic solids such as the nitride, oxide and halide insulators, and compound semiconductors. The four electron-pair bonds of C, Si, and Ge are formed by the four valence electrons of each host atom.

The four valence electrons are: $(2s)^2(2p)^2$ in diamond (carbon), $(3s)^2(3p)^2$ in Si, and $(4s)^2(4p)^2$ in Ge. From Table 162.2, we see that the radii of the radial charge density of these four valence electrons are four to five times larger than the next inner shell electrons. For example, the electron radii in isolated Si atom are: 0.904A for the two $(3s)^2$ electrons and 1.068A for the two $(3p)^2$ electrons while the next inner shell radii are 0.24A $(2s)^2$ and 0.20A $(2p)^2$. The interatomic spacing in a Si crystal is about 2.35A from x-ray diffraction data. Thus, only the outer shell electrons are expected to participate in the conduction of electricity by Si, just like their dominance on chemical reactions, since the inner shell electrons are so tightly bound to atomic nucleus. These outer-shell electrons are the very **valence electrons** which form the chemical or **valence bonds** and give the term, **valence electron bond model** in semiconductor physics. The inner shell electrons are known as the **core electrons**. The nucleus and the core electrons together are known as the **atomic core**, or **ion core**. A sketch of the electron charge distribution around the silicon nucleus, roughly to scale, is shown in Fig.171.1.

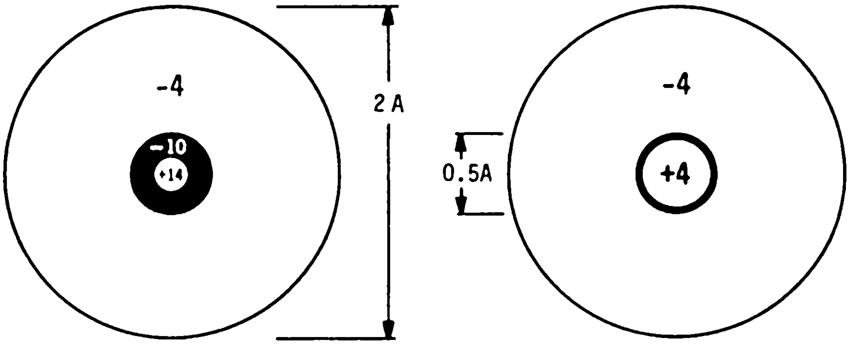


Fig.171.1 The electron charge distribution around a silicon nucleus with +14q.

The interatomic force that binds the atoms together in the III-V compound semiconductors, such as GaAs and GaP, is mainly from the electron-pair bonds. There is a slight ionic force because the valence electrons are not evenly shared by the two different adjacent host atoms (such as Ga and As). Their different core charge distributions give different valence electron radii. From Table 162.1, the valence electron radii are $(4s)^2=0.960A$, and $(4p)^1=1.254A$ for Ga; and $(4s)^2=0.826A$ and $(4p)^3=0.99A$ for As. Thus, the $(4p)^1$ gallium electron of 1.254A radius spends more time around the As atom in one of its unoccupied (4p) state than around the As atom. This electron charge transfer from Ga to As makes Ga more positive than As and gives GaAs a slight ionic character. This charge transfer can be represented by $Ga^{+\delta}As^{-\delta}$ where δ is a fraction of the electron charge transferred from Ga to As. In spite of the ionic character, the simple picture of the valence bond model is still very effective for explaining the electrical properties of compound semiconductors, such as the III-V compound semiconductors just discussed, the II-VI (ZnS, CdS, ...) compound semiconductors which have attracted recent and renewed interests, the I-VII halide (NaCl, NaF, KCl, KF, etc.) insulators, and the nitride (Si_3N_4) and oxide (SiO_2) insulators. The ionic character, known as **ionicity**, increases from the III-V to II-VI semiconductors. It increases further in the fully ionic I-VII insulators.

Three pictorial representations of the covalent bond or electron pair bond model are shown in Fig.171.2 on the following page. The larger circles in these figures represent the Si^{+4} atomic cores. The lattice or interatomic core spacing and the core diameter are drawn to scale for Si which is labeled in the inset on the right side of figure (a). The inset also shows the lobe of the sp^3 wavefunction or electron orbital from the adjacent core and the overlap of the two lobes form the electron-pair bond. Figure (a) is the Shockley model in which each rod or stick represents one electron and a missing rod represents a hole. A circle is placed at the missing rod or stick to make the presence of a hole more vivid. Figure (b) is the stick and ball or rod and ball model used by chemists. It is less desirable since it is less vivid to show the presence of a hole. Figure (c) is the dot model in which the electron pair bond is represented by two dots. A circle can be added to represent a hole at a missing dot. A circle in the space between the cores and outside of the bond regions is meaningless. It is obvious that the dot model in figure (c) is the most desirable model since it gives a vivid picture at one glance and it represents the electron by only one symbol, the dot.

Figure 171.2(a) also gives an illustration of an application of the bond diagram. A covalent or electron-pair bond is broken and one of the two electrons in the bond is released to the space between the bonds and the atomic cores, leaving a hole behind in the position where the bond electron was. An **electron-hole pair** is generated. The pair generation can be effected by light if the photon energy is greater than the bond energy, about 1.2eV in Si, requiring a photon wavelength shorter than $\lambda = 1.24/E = 1.24/1.2 = 1\mu m$ which is in the infrared range. When an electron-hole pair is generated, the solid is known as in an **excited state**. When all the valence electrons are at the lowest possible energy levels, the solid is known as at the **ground state**. Thus, optical generation of electron-hole pairs is sometimes known as optical excitation of the solid.

Bonds can also be broken or ruptured if the solid is at a finite temperature by being in contact with an ambient or heat sink which is at a finite temperature. In this case, bonds are broken and electron-hole pairs are generated, or shaken loose, by the **random thermal vibration** of the atomic nuclei. These random vibrations come from the collisions of the ambient gas molecules with surface Si atoms on the Si surface. The random collisions cause the surface atoms to vibrate randomly.

The vibration of the surface atoms then propagates into the interior of the solid and causes the interior atoms to vibrate randomly which shake loose some of the valence bond electrons. In vacuum, the surface atoms vibrate randomly owing to exposure to radiation or photons. There are many other useful mental pictures using the bond model such as the description of the effects of impurities on material and device properties given in chapters 2 and 3.

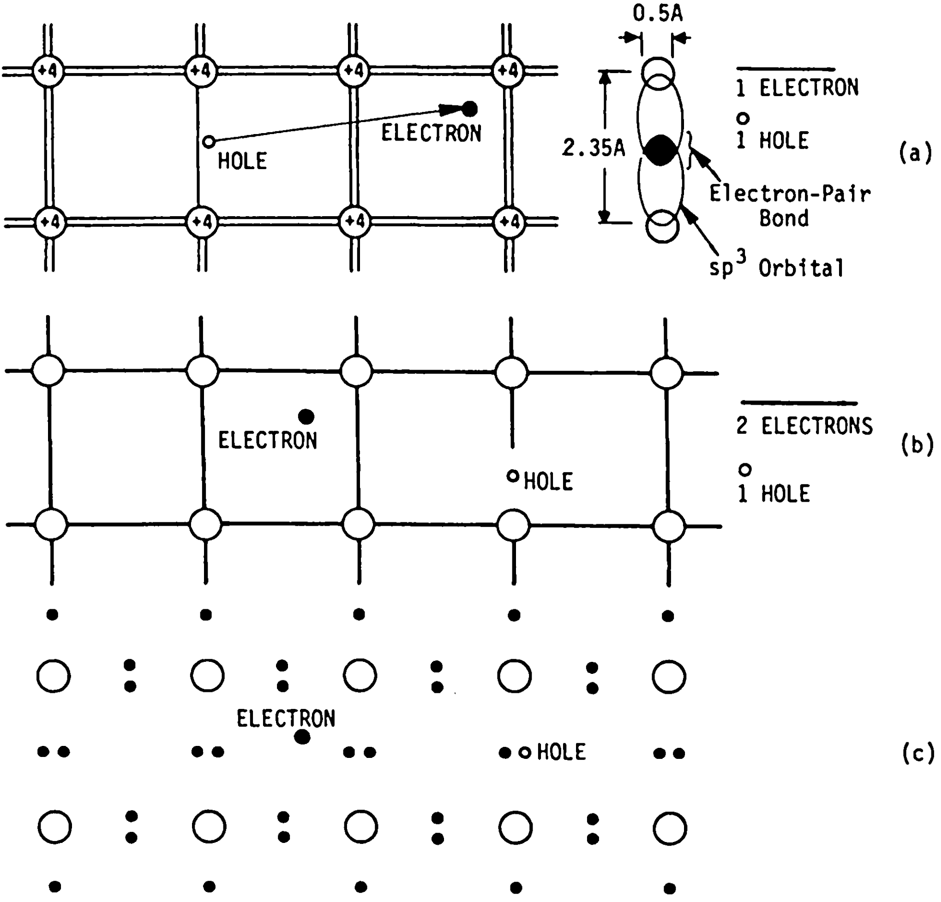


Fig.171.2 Valence electron bond models of silicon crystal. (Actual projected plane view is hexagonal - see Fig.132.2. Inter-Si core spacing - 2.53A.) Large circles are Si^{+4} ion core (diameter - 0.5A). Small circles are holes or missing bond electron. The inset on the right shows the lobe of the sp^3 wavefunction or orbital, two of which form a covalent electron-pair bond. (a) Shockley's model. Each rod or stick represents one electron. (b) Chemist's stick and ball model. Each rod or stick represents two electrons or one electron-pair bond. (c) Chemist's dot model. Each dot represents one electron. Two dot is an electron-pair bond.

172 The Band Model

The bond model is inadequate to provide a mathematical starting point from which to develop a **quantitative** theory of the properties of semiconductor since it does not give the spatial variation of the electric potential and the electric field. The **energy band model** or the **band model** must be used. The band model can be developed from the potential energy diagram of the electrons. The source of the potential energy felt or possessed by the test electron (as the test charge concept in electrostatics) is the Coulomb electrostatic force from all the ion cores and all other valence electrons in the solid. Thus, it is also known as the **one-electron energy band model**; it is the energy bands seen by the one test electron. However, the bond model will help us to formulate a good band model because it makes the physics and mathematics of the band theory more visual and easier understandable.

The starting point of developing the one-electron energy band model of solids is the potential energy diagram of an electron in the isolated neutral host atom. This is the reason for spending some time to study the energy levels and the wavefunctions of an isolated hydrogen atom. The atomic one-electron energy levels and wavefunctions are then used to build up the one-electron energy level diagrams in a crystal with many atoms and electrons.

Figures 172.1(a) to (c) show the step-by-step development of the one-electron energy bands in solid from the discrete energy levels in an isolated atom. The numerical energy level values are given for Si. The important energy levels are shown as continuous horizontal lines and labeled, such as E_C.

Figure (a) shows the energy levels of an isolated neutral Si atom calculated from the normalized values given in Table 162.1. Note that the valence shell levels are: $(3p)^2$ at 6.53eV below the vacuum level (VL=0eV) and $(3s)^2$ at 13.57eV below the vacuum level. The negative signs indicate that they are below the vacuum level while the Table 162.1 gives positive a sign since it is for the binding energy or the ionization potential.

Figure (b) shows how the energy levels of the isolated atom are shifted and broadened into bands of energy levels when many (neutral) Si atoms are moved closer to each other to form a crystal. The x-axis in figure (b), labeled a, is the nearest neighbor spacing 'a' of the Si atoms. 'a' is the hypothetical independent variable which decreases from $a=\infty$ for isolated atoms to $a_0=2.35A$ which is the equilibrium interatomic spacing between the nearest neighbor Si atoms in an Si crystal. This figure shows that as 'a' is decreased, the 3s and 3p valence electron energy levels are broadened into two bands of energy levels, known as **energy bands**, separated by an energy gap, E_G, [labels in figure (c)]in which there is no allowed energy levels. a_0 is the equilibrium interatomic spacing. The energy band above the energy gap is known as the **conduction band** in this Si example. The upper edge of the conduction band, labeled $E_{C'}$, is the vacuum reference level. The lower edge of the conduction, labeled E_C, is the upper edge of the energy gap. The energy band below the energy gap is known as the valence band. The upper edge of the valence band, labeled E_V, is the lower edge of the energy gap.

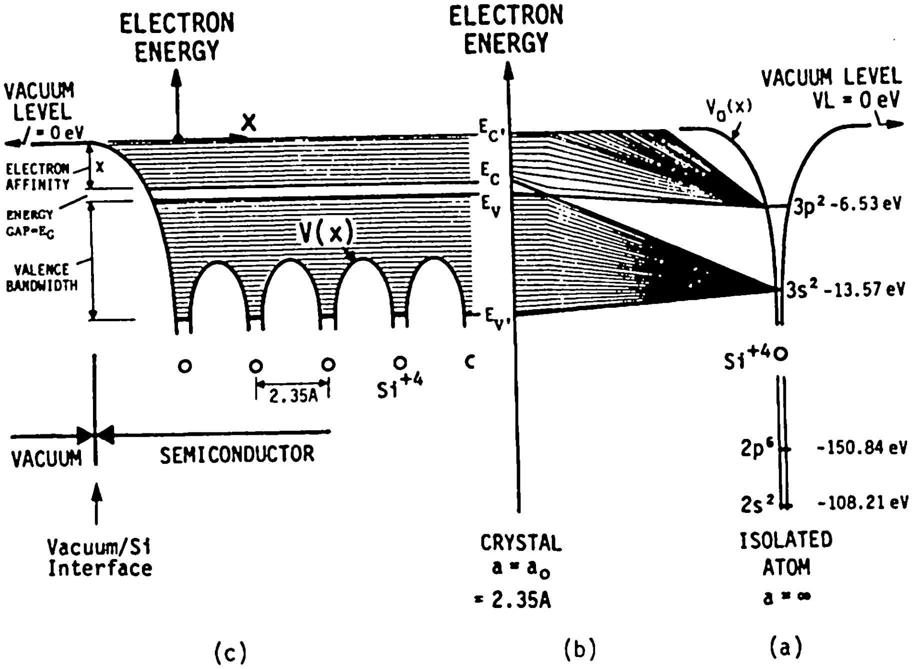

ELECTRON ENERGY
VACUUM LEVEL
= 0 eV
X
ELECTRON AFFINITY
χ
ENERGY GAP=E_G
VALENCE BANDWIDTH
V(x)
E_C'
E_C
E_V
E_V'
2.35A
Si+4
VACUUM
SEMICONDUCTOR
Vacuum/Si Interface
ELECTRON ENERGY
VACUUM LEVEL
VL = 0 eV
V_0(x)
3p² -6.53 eV
3s² -13.57 eV
Si+4
2p⁶ -150.84 eV
2s² -108.21 eV
CRYSTAL
a = a_0
= 2.35A
ISOLATED ATOM
a = ∞
(c)
(b)
(a)

Fig.172.1 (a) The atomic potential energy, $V_a(x)$, and energy levels of an electron in an isolated Si^{+4} ion. (b) The shift of the electron energy levels as many isolated Si^{+4} ions move together to form a silicon crystal at an equilibrium interatomic (inter-ionic) spacing of $a=a_0=2.35$A. (c) The crystalline potential energy, V(x), and the bands of energy levels of an electron in a silicon crystal, including details near the silicon surface exposed to vacuum.

It is clear that both the conduction and valence band energy levels come from the valence electron energy levels, $(3s)^2$ and $(3p)^2$, of the isolated Si atom. Theoretical solid-state and condensed matter physicists called both the conduction and valence bands, the valence band states for this reason. The separate terms, conduction and valence bands, were adopted and made popular by Bardeen, Brattain, and Shockley when they discovered and invented the transistor, in order to give a convenient and simple distinction while talking about electrons in the conduction band and holes in the valence band. The term **energy gap** has also been known as the forbidden energy gap, the forbidden gap, the band gap, and others. The last, band gap, is a misnomer; it is obscure since it suggests band and gap, although it is frequently used by engineers to denote the forbidden energy-gap. **Energy gap**, first adopted by John C. Slater and William Shockley, is the preferred term and it is used throughout this book.

Figure (a) also shows that the deep core energy levels are hardly broadened, such as $(2p)^6$ at -108.21eV and $(2s)^2$ at -150.84eV. The basic physics reasons for the shifting and banding of the energy levels will be described shortly.

The heavy curves, V(x), in the one-dimensional schematic diagram of figure (c) show the detailed potential energy variation of (or seen by) the test electron due to the Coulomb electrostatic force from Si^{+4} cores and the other valence electrons. Note that far away from the surface or inside the Si **bulk**, the potential energy is nearly periodic with the periodicity a_0(=2.35A for Si). The figure also shows the fine features of the potential variation near the solid surface and at the solid-vacuum interface. For example, the potential energy rises towards the vacuum level as the electron moves away from the Si surface into vacuum. For another example, the **potential barrier height** from the lower edge of the conduction band, E_C, to the vacuum level VL or $E_{C'}$, is known as the **electron affinity** and labeled χ. It is the binding energy of an electron by a neutral crystal and it is the solid-state analog of the very quantity used in isolated neutral atoms and molecules. The electron affinity is larger (or more negative) in solid than in isolated atom owing to the large number of atoms in the solid. For example, it is 4.018eV for Si crystal and 1.385eV for Si atom. The fine horizontal lines in figure (c) and the fine sloped lines in figure (b) are the energy levels or the allowed solutions of the many-electron and many-atom Schrödinger equation. Each line comes from one atom in the crystal.

The detailed energy band diagram given in atomic dimension (1cm≈2.35A) in Fig.172.1(c) is compressed in Fig.172.2 to practical transistor dimensions (1cm≈1000A) in order to use the energy band diagram to analyze transistor current-voltage characteristics. The closely spaced (allowed) energy levels in the conduction and valence bands are shown as horizontal lines. The numerical energy values are given for Si at room temperature. The electron affinity is χ_{Si}=4.02eV which is also the width of the conduction band, the energy gap is E_G(Si)=1.18eV, and the valence band width is about 12eV. The valence band width does not enter into device theory and hence is generally ignored in device studies. It is obvious that the more appropriate term is **band height** instead of band width but the traditional usage, **bandwidth**, will be adhered to. These numerical energy levels values and the widths of the energy bands and gaps are all determined experimentally. Theoretical calculations have not been sufficiently accurate to predict these energies to better than about 0.1eV. This has been improved to 0.03eV reported in 1990 research articles.

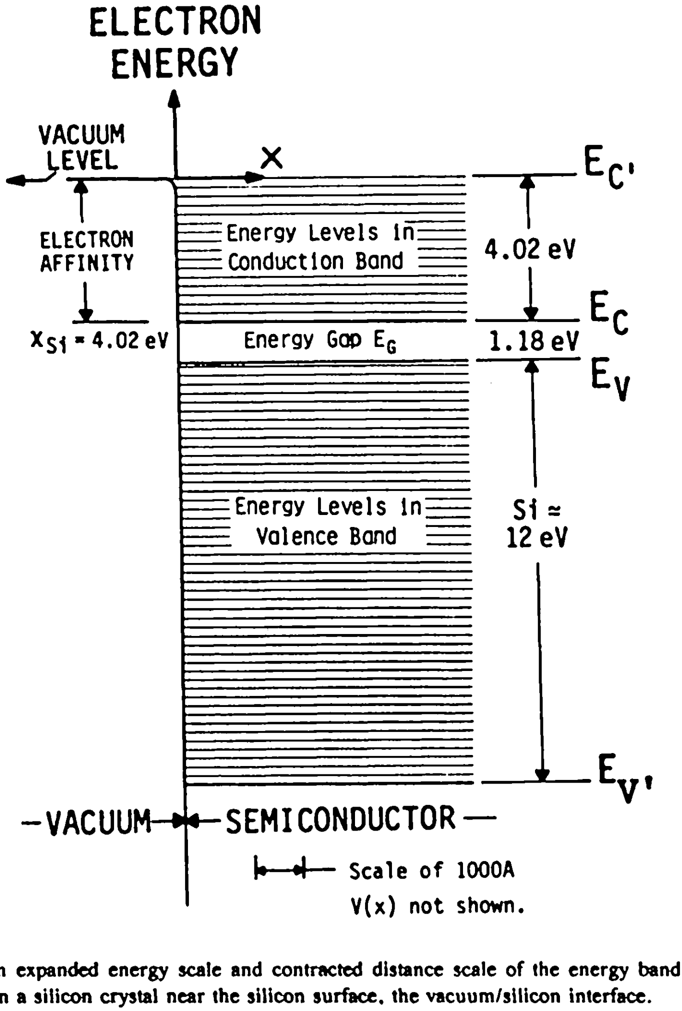

ELECTRON
ENERGY
VACUUM
LEVEL
X
ELECTRON
AFFINITY
χ_{Si} = 4.02 eV
Energy Levels in
Conduction Band
Energy Gap E_G
Energy Levels in
Valence Band
$E_{C'}$
4.02 eV
E_C
1.18 eV
E_V
Si ≈
12 eV
$E_{V'}$
-VACUUM
SEMICONDUCTOR —
Scale of 1000A
V(x) not shown.

Fig.172.2 An expanded energy scale and contracted distance scale of the energy band diagram of an electron in a silicon crystal near the silicon surface, the vacuum/silicon interface.

Some of the basic features and physics underlying the formation of the energy bands are described in the following numbered paragraphs. These concepts are useful in helping to enumerate the energy levels and energy states and to show how the states or levels are filled by electrons, discussed in the following section, 173. They are also helpful for formulating the mathematical theory of the energy band model in later sections, 18n.

(1) Many Atoms Give Many Energy Levels

There are many atoms in a solid, so that one expects many energy levels. Consider the 1s energy level of an isolated atom. When we put 10 ($N=10$) of these atoms on a line at a constant interatomic spacing 'a', we should have 10 1s-like energy levels. Since each of the 1s-like level has a 2-fold spin degeneracy, we would have 20 1s-like states in this linear or one-dimensional finite crystal. We use the term '1s-like' because the energy level and the wavefunction of the electrons in the crystal are not identical to those in an isolated atom and the 1s designation was used in a 1-electron isolated atom. In section 162, we have already seen a modification of the 1s hydrogen electronic levels and 1s wavefunctions when they are applied to a heavier and many-electron atom.

(2) Potential Energy of Electron is More Negative

The potential energy of a test electron in the crystal, $V(x)$ in Fig.172.1(c), is more negative than in an isolated atom, $V_a(x)$ in Fig.172.1(a). This lowering potential energy arises from the presence of the nearby atoms around an A-atom (for convenience of discussion, let us call this 'test atom' the A-atom) in a crystal, because the positive charge from the adjacent atomic cores are not completely screened by the negatively charged valence electrons. The uncompensated additional positive charges then lowers the potential energy of the test electron around the A-atom or any host atom in the crystal.

(3) Potential Energy of Electron is Periodic

The potential energy of the test electron, $V(x)$, is periodic if we consider an infinitely large crystal as the ideal model. For regular crystals, the potential energy is almost periodic since there are so many atomic layers even in the thin surface layer of a transistor (1 micron or 10^4A thick Si layer has about 5000 atomic layers of Si). Only at the vacuum/solid interface or the physical surface, is there a disruption of the periodicity as indicated in Fig.172.1(c). Such a disruption gives bound states known as surface traps which are recombination-generation-trapping sites of electrons and holes. Traps degrade transistor performance because the electrons and holes disappear at the traps via recombination and because electrons and holes are randomly generated by the traps which give leakage current and noise.

(4) Electron Energy Levels are Shifted Upwards and Downwards

The electron energy level positions in an isolated atom are shifted when many atoms are moved closer together to form a crystal. The shifts can be either upwards or downwards depending on the spatial symmetry of the electron wavefunction or probability distribution. Figure 172.1(b) shows upwards and downwards shifts.

(5) Energy Levels are Clustered into Energy Bands

The energy levels from one atomic level, say 3s, described in paragraph (1) and (4) above, no longer have just one discrete value. If they all had only one value, we would have a 3s level with 10-fold (N=10 Si atoms) configuration degeneracy or 20-fold total degeneracy (counting the 2-fold spin degeneracy) in the 10-atom linear chain. Instead, the energy levels are shifted to cover a range or a band of energies which gave the term **energy band**. A simple way to look at this banding of energy levels is the use of the tunneling model and Heisenberg's uncertainty principle between energy and localization time. When the two atoms are nearby, such as in a crystal, the potential energy barrier between the two nuclei is thin and not-very-high. [See the rounded hills between adjacent atomic cores in Fig.172.1(c).] Thus, the electron around the A-nucleus will have a large probability to appear at or around the adjacent B-nuclei. The concept of tunneling and probability wavefunction can be used to make a quantitative estimate of the localization time and the energy uncertainty or the width of the energy band. The uncertainty principle gives $\Delta E \Delta t \propto h$, thus, $\Delta E \propto h/\Delta t$. $(1/\Delta t)$ is proportional to the tunneling rate. Thus, closer nuclei reduces the height and thickness of the potential wall separating the atoms and increases the tunneling rate which gives large energy spread or large energy band width.

(6) Electrons are No Longer Localized

Electrons in a crystal can no longer be localized to a single atomic nucleus, while in an isolated atom it is **truly localized** in all three space dimensions or truly **bound** to the positively charged point nucleus. The electrons in the crystal are actually distributed over the entire crystal although their wavefunctions or probability densities may peak around each nuclear position. The uncertainty principle between position and momentum may be used to estimate the position delocalization or the spread of the electron charge's probability density function.

173 Filling the Energy Band Levels by Electrons

In the preceding section, the formation of the energy band model was described qualitatively. The next question concerns filling of the energy levels by electrons. The answer is crucial for developing the concept of the conduction electrons and holes in the energy band model of semiconductors.

Figures 172.1(a) and (b) illustrated the shift of the atomic energy levels when isolated silicon atoms are brought together to form a Si crystal. An expanded view is given in Figs.173.1(a) and (b) which also illustrate how the band of allowed energies are filled by the electrons which we shall explain.

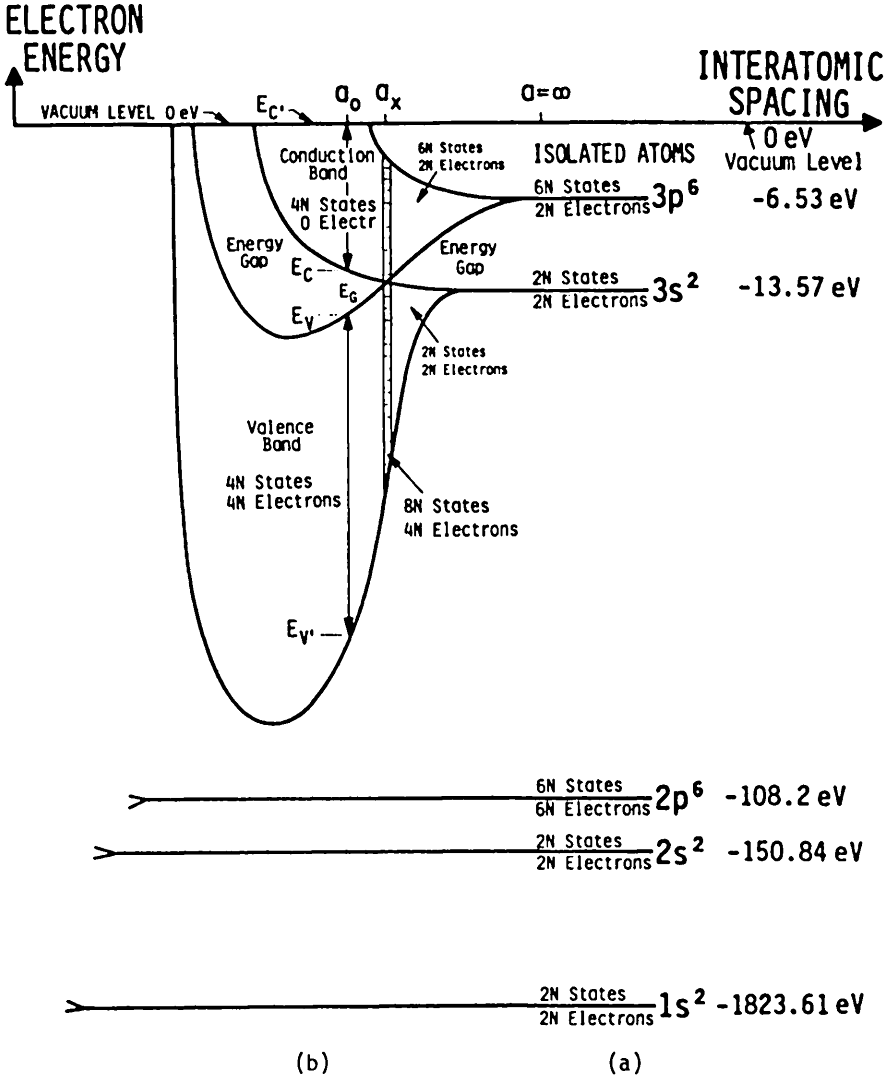

ELECTRON ENERGY
INTERATOMIC SPACING
VACUUM LEVEL 0 eV
$E_{C'}$
a_0
a_x
$a=\infty$
0 eV Vacuum Level
Conduction Band
4N States
0 Electr
6N States
2N Electrons
ISOLATED ATOMS
6N States
2N Electrons
3p⁶
-6.53 eV
Energy Gap
E_C
E_G
E_V
Energy Gap
2N States
2N Electrons
3s²
-13.57 eV
2N States
2N Electrons
Valence Band
4N States
4N Electrons
8N States
4N Electrons
$E_{V'}$
6N States
6N Electrons
2p⁶
-108.2 eV
2N States
2N Electrons
2s²
-150.84 eV
2N States
2N Electrons
1s²
-1823.61 eV
(b)
(a)

Fig.173.1 Electron energy levels and their occupation. (a) In N isolated Si atoms. (b) As a function of the interatomic spacing in Si. The energy axis is nonlinear. (c) The crossing and non-crossing valence bands for N=4 and 8 electrons. [Part (c) is on the next page.]

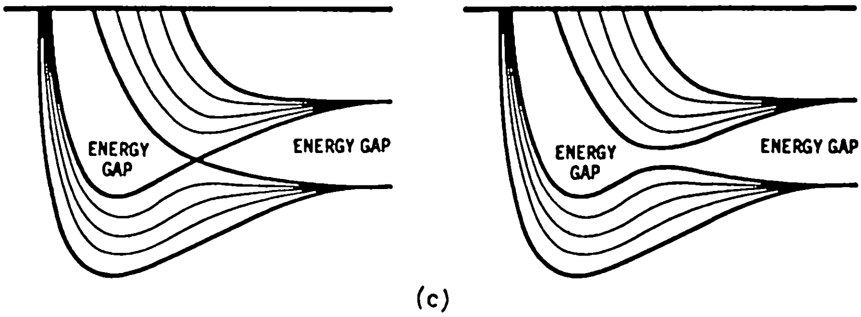


Fig.173.1 Electron energy levels and their occupation. (a) In N isolated Si atoms. (b) As a function of the interatomic spacing in Si. The energy axis is nonlinear. (c) The crossing and non-crossing valence bands for N=4 and 8 electrons. [Parts (a) and (b) are on the previous page.]

Consider an Si crystal composed of N Si atoms (let N=10). Then, there are NZ=140 electrons (Z=+14 for Si). How are the energy levels of the 10 Si atoms filled by the 140 electrons? According to Pauli's exclusion principal, no more than 2 electrons can have the same energy, wavefunction and spin. Thus, if two electrons have the same spatial configuration or probability distribution, then they must have opposite spin. Thus, each energy level can be occupied by two electrons provided they have opposite spin.

To determine how the electrons will occupy the energy levels, we shall consider the situation at absolute zero temperature in order to freeze the nuclei at their lattice points so that they contribute no kinetic energy. In chapter 2, we shall remove this restriction in order to include vibrating nuclei or atomic cores. At equilibrium, all the electrons are trying to move to the lowest energy. But Pauli's exclusion principle prevents more than two electrons from occupying one energy level.

Consider first the ten (N=10) isolated or widely separated Si atoms shown in Fig.173.1(a) which has a total of NZ=10x14=140 electrons. There are 20 (1s) states from the 10 Si atoms. The 20 states will be occupied by 20 electrons.

Similarly, the 10 2s levels are occupied by 20 of the remaining 120 electrons. The 30 3p levels are occupied by 60 of remaining 100 electrons, leaving the 40 valence electrons unaccounted for thus far.

How do these 40 remaining electrons occupy the valence energy levels? This is shown in the upper part of Fig.173.1(a). When the Si atoms are far apart or isolated, half or 2N(=20) of these 4N(=40) valence electrons will occupy the

2N(=20) 3s states. The remaining 2N(=20) electrons will occupy 2N(=20) of the 6N(=60) 3p states.

From the above enumeration, we conclude that the 1s, 2s, 2p and 3s levels of the isolated Si atom are all completely filled, and the 3p levels are partially filled by electrons.

When the atoms are brought together to form a crystal, the electronic energy levels of isolated atom are shifted due to the presence of adjacent atoms as discussed in Fig.172.1(b). The effects are the largest for the largest orbits such as those of the 3s and 3p valence electrons. This is illustrated in Fig.173.1(b) for Si. The forces due to the neighboring atoms perturb or shift the 3p energy levels from their original isolated-atom positions and lift or remove the configuration degeneracy. Spin degeneracy is unaffected by the Coulomb electrostatic force, only by magnetic force.

As the assumed atomic spacing decreases, Fig.173.1(b) shows that the shift of the 3p and 3s energy levels increases and the levels broaden into bands of allowed energies. At some interatomic spacing, the bottom level of the 3p band and the upper level of the 3s band will cross each other. This interatomic spacing is labeled a_x in Fig.173.1(b). Around this crossing, the energy levels will regroup from different mixing or linear combination of the 3s and 3p wavefunctions to give two possible distributions of allowed energy levels as a function of energy. One distribution does not have an energy gap which is shown in Fig.173.1(b) and the left part of figure (c). The other has an energy gap which is shown in Fig.173.1(c). In either case, we can no longer talk about the 3p-like and 3s-like energy levels of electrons in the upper and lower bands since the wavefunctions near the band edges are linear combinations of the 3s-like and 3p-like wavefunctions. Thus, the wavefunctions near E_V and E_C have both the 3s-like and 3p-like components. In the gapless case, at a_x shown in Fig.173.1(b), the combined band has 4N=40 levels (N from 3s and 3N from 3p) and 8N=80 states but only 4N=40 electrons (2N of 3s electrons and 2N of 3p electrons). Thus, only the lower half of the energy levels or electron states in this one gapless band are occupied by electrons in the band. This is a metal-like hypothetical case whose conduction electron density is equal to the valence electron density. This precise interatomic spacing is hard to maintain since the random vibration of the nuclei will cause the spacing to fluctuate.

As the assumed atomic spacing decreases further from the energy-level-crossing value of a_x, an energy gap again appears in the crossing case of Fig.173.1(b) and the energy gap widens in the non-crossing case of Fig.173.1(c). The 8N=80 states split equally between the upper band and the lower band. Since the electrons will seek and reside at the lowest possible energy states, the 4N=40 states of the lower band will be just completely filled by the 4N=40 valence electrons while the 4N=40 states of the upper band will be entirely empty or unoccupied by electrons since there are no electrons left: - all of them have been accounted for at the energy levels of the lower band. This is the situation that makes a semiconductor an insulator - there are no electrons in the conduction band and no holes or unoccupied states in the valence band. If the energy gap E_G is similar to or smaller than the photon energy of the visible and infrared light, about 0.1 to 3 eV, it is a semiconductor since some bonds are broken by photons and electrons and holes are generated. If the energy gap E_G is several times larger, greater than about 3 eV, it is an insulator since there are no electrons in the conduction band nor holes or unoccupied states in the valence band. The demarcation line distinguishing a semiconductor from an insulator is not sharp and highly dependent on temperature.

In summary, from the band or energy level filling picture just presented we note that the lower valence band is completely filled by electrons and the upper valence band is completely empty in Si. They are separated by an energy gap in Si, Ge as well as by other semiconductor and insulator crystals.

The energy band vs. position is redrawn in Fig.173.2(a) for a crystal containing eleven (N=11) Si atoms and 44 valence electrons in order to illustrate the generation of an electron-hole pair by light. Photon moves one electron into the conduction band and leaves one hole in the valence band. An electron-hole pair is generated. This pair generation process was demonstrated by the bond model given in Fig.171.2(a). The two examples, Fig.171.2(a) and Fig.173.2(a), show the **one-to-one correspondence between the bond and the band models**: the electron in the conduction band is the electron in the empty space between bonds and cores in the bond model. The hole in the valence band is the unoccupied bond in the bond model.

Figure 173.2(a) shows that it is already quite tedious to draw out all the energy levels in a band diagram for the Si crystal containing only eleven (N=11) Si atoms. In a crystal of $1cm^3$, there are about $N=10^{22}$ atoms. An energy band diagram with 10^{23} lines or energy levels would have completely dark conduction and valence bands due to the overlapping lines. In order to use the energy band diagram to analyze the characteristics of diodes and transistors, the simplified diagram shown in Fig.173.2(b) is commonly used. The energy level lines are all removed except those of the band edges. The valence electron dot symbols are also removed from the valence band. Instead, a missing valence electron or an unoccupied energy level in the valence band is now represented by a circle, a **hole**. An electron in the conduction band is still represented by a dot. This will be the energy band diagram to be used in device analysis of the following chapters. It is known as the **energy-distance** or E-x energy band diagram. It is the potential energy band diagram of the one test electron in the physical or **position space**. There is a complement energy diagram in the **momentum** (velocity) or **wave-number space**, known as the E-k or **energy-wave number** diagram. The theory of the E-k diagram will be developed in sections 18n on the mathematical theory of the energy bands.

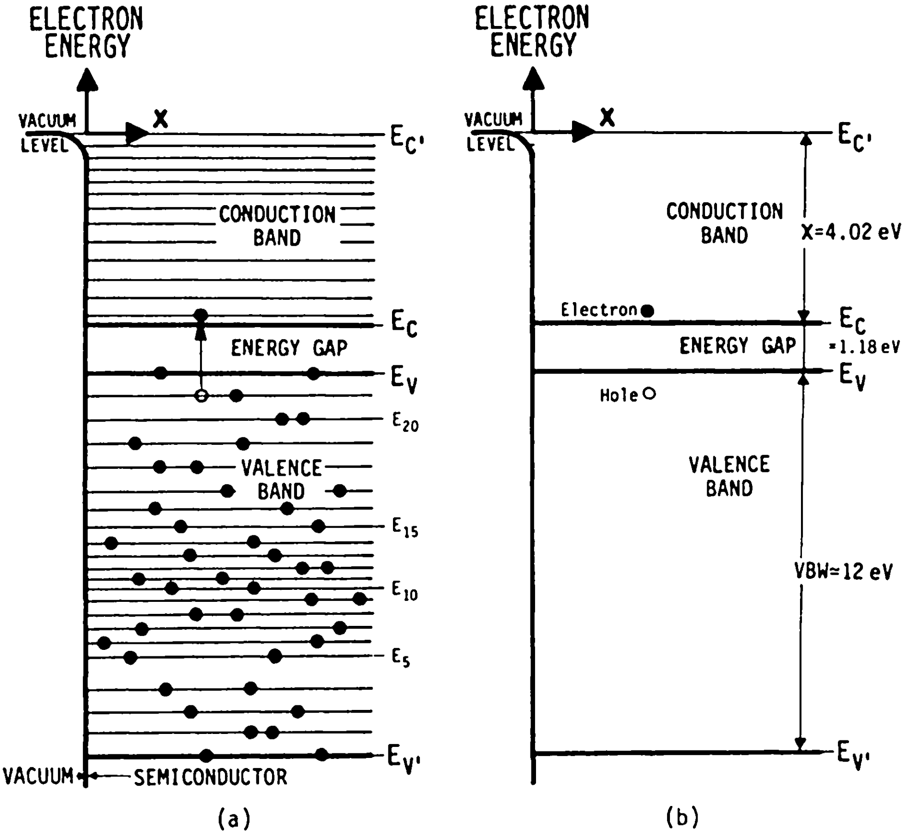

ELECTRON ENERGY
VACUUM LEVEL
X
$E_{C'}$
CONDUCTION BAND
E_C
ENERGY GAP
E_V
E_{20}
VALENCE BAND
E_{15}
E_{10}
E_5
$E_{V'}$
VACUUM
SEMICONDUCTOR
(a)
ELECTRON ENERGY
VACUUM LEVEL
X
$E_{C'}$
CONDUCTION BAND
X=4.02 eV
Electron ●
E_C
ENERGY GAP
=1.18 eV
E_V
Hole ○
VALENCE BAND
VBW≈12 eV
$E_{V'}$
(b)

Fig.173.2 (a) A sketch of the Si energy band diagram $N=11$. (b) A simplified energy band diagram, the **E-x diagram**, to be used for device analysis.

180 ELEMENTARY DERIVATIONS OF THE ENERGY BAND MODELS

There are two fundamental methods to the calculation of the electron energy levels or energy bands in semiconductors and solids. They are fundamental both in physics and in mathematical technique. Several frequently used names of these two methods are listed below.

(1) **Nearly-Free Electron Model** (also known as)

Loosely bound electron model (invented here)
Plane wave (PW) method
Fourier series expansion method
Brillouin method

(2) **Tight-Binding Model** (also known as)

Tightly bound electron model
Linear combination of atomic orbitals (LCAO) method
Wannier series expansion method
Bloch method

The term 'bound' used here is literal and not mathematical nor quantum mechanical. In quantum mechanics, we recall that a bound state is a very precisely defined electron state whose wavefunction is bound or decays in all three space directions in order to give a truly localized state. The magnitude of the square of the wavefunction, when integrated over all space, is bound or has a finite value. To conform with the probabilistic interpretation of the square of the wavefunction, this finite value can be set to unity by a proper choice of the coefficient multiplying the wavefunction. Thus, 'tightly bound electron model' is not used to avoid confusion.

As the names suggest, method (1) is particularly good for high energy electrons and method (2) for low energy electrons. These are illustrated by the two electrons labeled by circled numerals 1 and 2 in the energy diagram given in Fig.180.1. This diagram is an abbreviated enlargement of Figs.172.1(c) and Fig.173.2(a). In Fig.180.1, only the two demonstration electrons are explicitly shown as dots and labeled by circled numerals 1 and 2. All the other electrons in the valence and conduction bands are omitted. We shall give two simple mathematical analyses, one for each of the two models, to show how the energy bands are formed.

These two energy band calculation methods will also give us a second way of illustrating the energy bands. The first way was the one we have used so far: **Energy vs Distance (the E-x diagram)** shown in Figs.172.1(c), 172.2, 173.2(a), 173.2(b) and 180.1. The use of the term 'E-x' diagram was first made by this author in 1961 to simplify the teaching of semiconductor physics to undergraduate juniors. The second way is the **Energy vs Momentum or Wavenumber** (the **E-k**) diagram which shall be derived and illustrated in the next two subsections, 181 and 182.

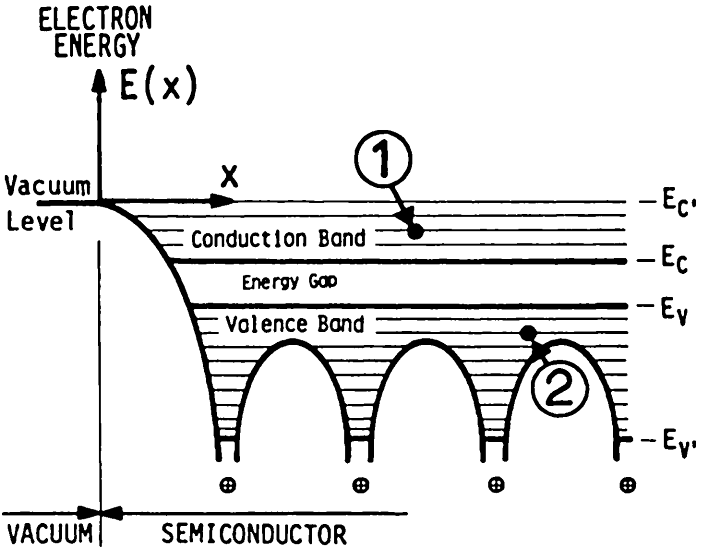


Fig.180.1 The energy band diagram of a solid to illustrate a high-energy nearly-free electron (circled numeral 1) and a low-energy tightly-bound electron (circled numeral 2).

181 The Nearly Free Electron Model

As the name suggests, this model starts out by ignoring the potential from the ions or atomic cores of the crystal. In the following discussions, we shall frequently omit the words 'electron' and 'energy' in the term 'electron potential energy'. Potential and potential energy have the same numerical value if energy is given in electron-volt unit. The potential due to an isolated and positively charged atomic ion core is similar to the Coulomb potential shown in Fig.141.1. It was used to sketch the electron's potential energy arising from an isolated Si^{+4} in Fig.172.1(a). If the atomic potentials, $V_a(r)$, are added, the **periodic potential** V(x) of electron in a crystal is obtained, such as V(x) in Fig.172.1(c). When we ignore this periodic potential, we have a spatially constant potential, V(x)=constant, except near the surface which is also ignored by considering the middle part of a large crystal which is far away from the crystal surfaces. The solutions of the electron wavefunctions of the Schrödinger equation in a constant potential are plane waves, exp(ikx). The energy-momentum relationship of the plane waves is $E=p^2/2m=\hbar^2k^2/2m$. Thus, if we plot energy, E, versus wave number, k, we have a continuous parabola given by the thin curve shown in Fig.181.1(b) and the energy levels are continuous.

If we include the steeply rising potential near the two surfaces of a slice of crystal (left surface is shown in Fig.180.1 while right surface is off the page), then we have a box or square well potential problem shown in Fig.181.1(a). When the box is made very large to model a semiconductor crystal in a practical situation, such as 1 cm or 1 mm thick silicon slice, we again have the plane wave solution. So a box potential such as that shown in Fig.181.1(a) is useless although it has been used by many textbooks as the crystal model to compute the volume density of the electron states. **The box model and square well potential used by many textbooks for a crystal are not only mathematically more complex but also conceptually deficient, misleading, erroneous, and irrelevant.**

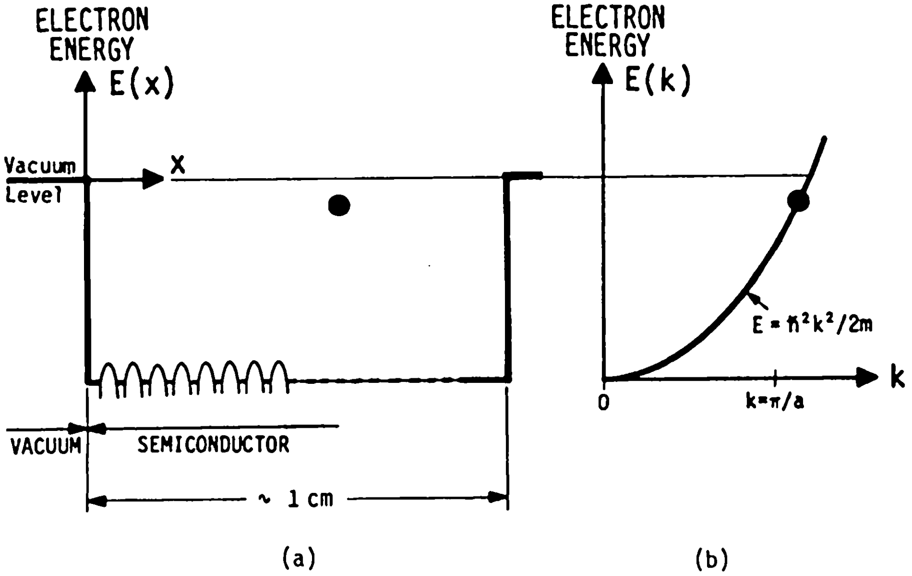


Fig.181.1 (a) Square-well potential of the box model of a crystal. (b) The free-electron energy wave-number parabola.

To model a semiconductor or a crystalline solid correctly, we must include the periodic potential, V(x), from the host atoms. Mathematically, this periodic potential is introduced by increasing its amplitude from zero as suggested by the small-amplitude periodic potential in Fig.181.1(a) to a the value of a real crystal (about 10 eV below the vacuum level) shown in Fig.181.2(a). The finite-amplitude of the periodic potential causes the continuous and parabolic E-k diagram of a free electron to break up into bands of allowed ranges of energy as indicated by the heavy curve in Fig.181.2(b). The two adjacent allowed ranges of energy are separated by a energy range where there are no travelling wave solutions. This is known as the energy gap and sometimes called the forbidden energy band or the misnomer, band gap. Figure 181.2(b) shows only the lowest energy gap. There are many energy gaps at higher energies because the periodic potential V(x) has many higher space harmonic components.

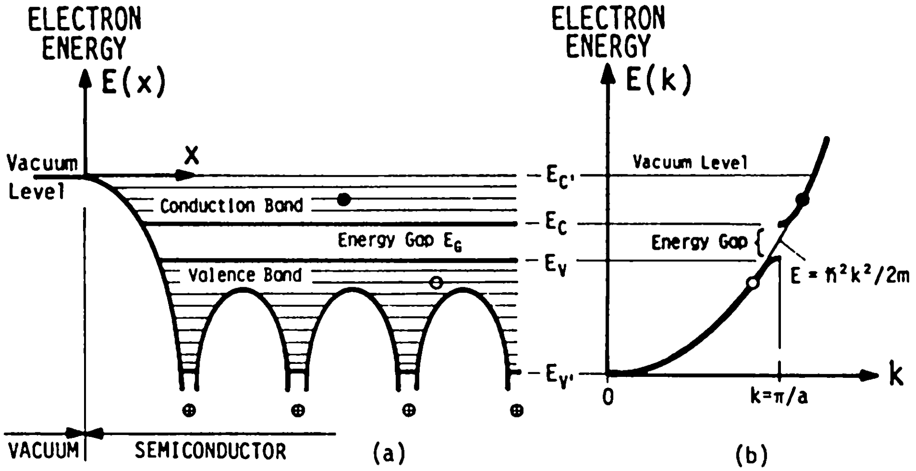


Fig.181.2 (a) The real periodic potential of an electron in a crystal, V(x). (b) The energy-wave number diagram showing the free electron parabolic is broken up into two parts by the presence of the periodic potential, V(x).

The mathematical analysis used in the nearly free electron model to derive the energy-wave number relationship or the E-k energy band diagram is an exercise of Fourier analysis. It is a good elementary example on using the Fourier series expansion method to solve a position dependent differential equation. Because the potential function, V(x) of the crystal, in the Schrödinger equation is periodic with the periodicity of the crystal lattice, a, it can be expanded in Fourier series. We shall use the one-dimensional crystal as an example to illustrate the mathematics and physics that are used to obtain the approximate analytical solutions. Because V(x) is periodic in space, it can be expanded in Fourier series. This expansion is

$$V(x) = \sum_{\nu} V_{\nu}\exp(iK_{\nu}x) \qquad (181.1)$$

$$V_{\nu} = \frac{1}{a}\int_{0}^{a} V(x)\exp(-K_{\nu}x)dx \qquad (181.2)$$

where $\nu=0,\pm1,\pm2,\pm3...$ and $K_{\nu}=2\pi\nu/a$. The V_{ν} are the Fourier series expansion coefficients of the periodic potential, V(x). They can be obtained by the standard inversion formula given by (181.2). The Fourier series expansion given above is the expansion in space harmonics. It is analogous to the more familiar time

harmonics used in the Fourier analysis of devices and circuits subjected to time-varying electrical-mechanical-optical forces.

The general solution of the Schrödinger equation with a periodic potential V(x) is given by $\psi(x)=\exp(ikx)u(x)$. It is a plane wave, exp(ikx), modulated by a function u(x). It can be shown that **modulus** u(x) is a periodic function with the periodicity of the crystal lattice, i.e., it is **cell periodic** or it repeats itself from one unit cell to the next unit cell. The proof of u(x) being cell periodic is known as the **Bloch Theorem** in solid state physics and Floquet Theorem in the theory of differential equation with periodic coefficients. We will not give this proof because the algebra is somewhat lengthy, although straightforward. However, we will make use of the consequence of the Bloch Theorem, that is: u(x) is cell periodic or $u(x) = u(x+a) = u(x+na)$ where n is a positive or negative integer. Since u(x) is periodic, it can again be expanded in a Fourier series just like the periodic potential V(x). Before writing down the expansion, we should note that the wavefunction written as a product of a plane wave, exp(ikx), and a periodic modulation, u(x), is an entirely reasonable general solution since if the periodic potential V(x) is very small or is made to approach zero, then we would get a plane wave solution, exp(ikx), and u(x) would be a constant. Thus, **u(x) represents the influence of the periodic potential, V(x), on the electron motion**. Because of the presence of the atoms or atomic cores situated at the lattice points in a crystal, V(x) is periodic instead of constant and u(x) is also periodic instead of constant. The **amplitude modulation function** or modulus, u(x), distinguishes the electronic motion in a solid from vacuum, in a crystalline solid from a non-crystalline solid, and in different materials such as molecules, solids, liquids and gas.

Next, expand the periodic function, u(x), by a Fourier series

$$u(x) = \sum_{n} C_n \exp(iK_n x) \qquad (181.3)$$

where C_n are the expansion coefficients, $n=0,\pm1,\pm2,\ldots$ and $K_n=2\pi n/a$. K_n is known as the **translation vector** (3-d) **of the reciprocal lattice** whose end points form an infinite periodic array of points in the k-space or reciprocal space. A more consistent notation would be U_n instead of C_n to emphasize that u(x) and U_n are Fourier transforms of each other just like the pair V(x) and V_ν. Substituting this expansion into $\psi(x)=\exp(ikx)u(x)$ yields

$$\psi(x) = \sum_{n} C_n \exp[i(k+K_n)x]. \qquad (181.4)$$

The Fourier expansion of $\psi(x)$ and V(x) is now substituted into the Schrödinger equation (to be abbreviated by SE in the following discussion)

$$-(\hbar^2/2m)d^2\psi(x)/dx^2 + V(x)\psi(x) = E\psi(x). \qquad (181.5)$$

These two substitutions convert the Schrödinger differential equation into a set of simultaneous algebraic equations of the unknown coefficients, C_n. [Note that V_ν are already known since V(x) is known or assumed known for a given crystal so that V_ν can be calculated using (181.2).]

There is an infinite number of simultaneous algebraic equations since there is an infinite number of C_n's (n=0 and all positive and negative integers, ad infinitas). Thus, some approximations must be used to compute the C_n's since we cannot invert an infinitely large matrix. The physics of the problem is used next to help to truncate the matrix and to make the algebraic approximation. Since we are looking at the nearly free electrons, this means that the influence of the periodic potential, V(x), is small. Thus, only the largest expansion coefficients of V(x) are likely to be important which are V_0, V_{+1} and V_{-1}. We also note that V(x) must be a real function of position. Thus, V_{-1} must be equal to V_{+1} if the origin of the coordinate is chosen such that V(x) is an even function or V(x)=V(-x). Then, the three-term approximation of V(x) is a cosine function of x with wave number $K_1 = 2\pi/a$ or wavelength 'a'. It is

$$\begin{aligned} V(x) &= V_0 + V_1\exp(iK_1x) + V_{-1}\exp(iK_{-1}x) \\ &= V_0 + 2V_1\cos(K_1x). \qquad (181.6) \end{aligned}$$

From the same reasoning given above for V(x), we would also expect only the lowest three terms of the Fourier expansion of u(x) to be important. Thus,

$$\psi(x) = [C_0 + C_1\exp(iK_1x) + C_{-1}\exp(iK_{-1}x)]\exp(ikx) \qquad (181.7)$$

By truncating the Fourier series expansion of V(x) and u(x), we have reduced the set of infinitely large number of simultaneous algebraic equations for C_n to a set of three equations with three unknowns, C_0, C_1 and C_{-1}. The number of equations and unknowns can be further reduced by one if we consider a region of solution where only either C_1 or C_{-1} dominates, then we will have only C_0 and C_1 or C_0 and C_{-1}. Let us find the solutions where only C_0 and C_{-1} dominate, then we have

$$\psi(x) = [C_0 + C_{-1}\exp(iK_{-1}x)]\exp(ikx). \qquad (181.8)$$

(181.8) and (181.6) can then be substituted into the SE, (181.5). It can then be inverted to give the 2 algebraic equations for C_0 and C_{-1}. The inversion is accomplished by multiplying the SE by exp(-ikx) and integrating over the interval x=0 to x=a. This gives

$$C_{-1}V_1 = [(\hbar^2k^2/2m) - E + V_0]C_0. \qquad (181.9)$$

Similarly, the SE is multiplied by $\exp[-i(k+K_{-1})x]$ and integrated over x=0 to x=a to give

$$C_0V_{-1} = [\hbar^2(k+K_{-1})^2/2m) - E + V_0]C_{-1}. \qquad (181.10)$$

Nontrivial or nonzero solutions of C_0 and C_{-1} requires that the ratio C_{-1}/C_0 from (181.9) and (181.10) are equal, i.e.,

$$C_{-1}/C_0 = [(\hbar^2k^2/2m)-E+V_0]/V = V_{-1}/\{[\hbar^2(k+K_{-1})^2/2m]-E+V_0\} \tag{181.11}$$

or

$$[E-V_0-(\hbar^2k^2/2m)]\cdot[E-V_0-\hbar^2(k+K_{-1})^2/2m] = V_1\cdot V_{-1} = |V_1|^2. \tag{181.12}$$

This quadratic equation in k can be plotted to give the E-k diagram which has an energy gap. A sketch was given in Fig.181.2(b).

The pair of equations for C_0 and C_{-1} given by (181.9) and (181.10) can be also written in the form of a 2x2 **homogeneous matrix equation** $[H_{ij} - E\delta_{ij}][C_j]=0$ where i and j are 0 and -1. It has only a non-trivial solution (i.e. $C_i \neq 0$) when its determinant is zero, $|H_{ij}-E\delta_{ij}|=0$. This gives the same result as (181.12).

We note two important features from (181.12) without having to make a detailed numerical analysis. (1) The energy E is shifted by an amount V_0. V_0 is the space average of the periodic potential V(x) in analogy to the time average or d.c. value, I, of a periodic function i(t). For a periodic potential such as that shown in Fig.181.1 (without the surface region) V_0 is negative or below the vacuum level (VL=0). (2) We see that at some values of k, the two square bracketed terms in (181.12) are equal. This equality occurs when $k^2 = (k+K_{-1})^2$ or when

$$k = \pm (k + K_{-1}) = -(k + K_{-1})$$

or

$$k = -K_{-1}/2 = \pi/a. \tag{181.13}$$

Physically, this is the condition of **Bragg reflection**. At this condition, the wavelength of the electron wave (given by $\lambda=2\pi/k$) is equal to twice the lattice spacing, 2a, or the electron wave is phase-shifted by $ka=\pi$ or 180° after traveling through a distance equal to one lattice spacing, a. At this wavelength, the electron can have two possible energies. This is demonstrated by substituting $k=\pi/a$ or $k=-\pi/a$ into (181.12) which gives

$$[E - V_0 + \hbar^2(\pi/a)^2/2m]^2 = |V_1|^2$$

or

$$E_\pm - V_0 = \hbar^2(\pi/a)^2/2m \pm |V_1|. \tag{181.14}$$

Thus, at $k=\pi/a$, we have two possible allowed energies. They are separated by an energy gap or a range of energy in which the Schrödinger equation has no wave-like solutions. This is the **energy gap** shown in Fig.181.2(b).

The size of the energy gap, E_G, is equal to $2|V_1|$ as indicated by (181.14) which gives $E_G \equiv E_+ - E_- = 2|V_1|$. This is an important result. It shows that **the energy gap is equal to twice the Fourier series expansion coefficient of the periodic potential.** It gives a simple physical meaning to the Fourier expansion coefficients of the periodic potential of the crystal. It also shows that there are many energy gaps since V(x) has many space-harmonic components. We only retained one in the above example.

Another useful illustration is the shape of the wavefunction at the energy gap. There are two wavefunctions at the energy gap, $k=\pi/a$, because the electron can have one of the two possible total energies, E_+ or E_-. From the condition that the two bracketed terms in (181.12) are both equal to $|V_1|$ or $-|V_1|$ which gave $E_G=2|V_1|$, (181.11) tells us that either $C_0=+C_{-1}$ or $C_0=-C_{-1}$. Obviously, one is for E_+ and the other is for E_-. The correct choice has proven elusive for generations! (See problem P181.8). But it is trivial if one remembers that E is the total energy or the average potential energy relative to the free-electron energy at $k=\pi/a$, $V_0+\hbar^2k^2/2m=V_0+h^2/2ma_2$. A simple illustration as follow can give the physics and the solution quickly. Using $C_0=\pm C_{-1}$ and $k=\pi/a$, (181.8) becomes

$$\psi_\pm(x) = C_0[1 \pm \exp(-i2\pi x/a)]\exp(i\pi x/a),$$

giving

$$\psi_-(x) = i2C_0\sin(\pi x/a)$$

and

$$\psi_+(x) = 2C_0\cos(\pi x/a). \qquad (181.15)$$

These are standing waves instead of traveling waves such as exp(ikx).

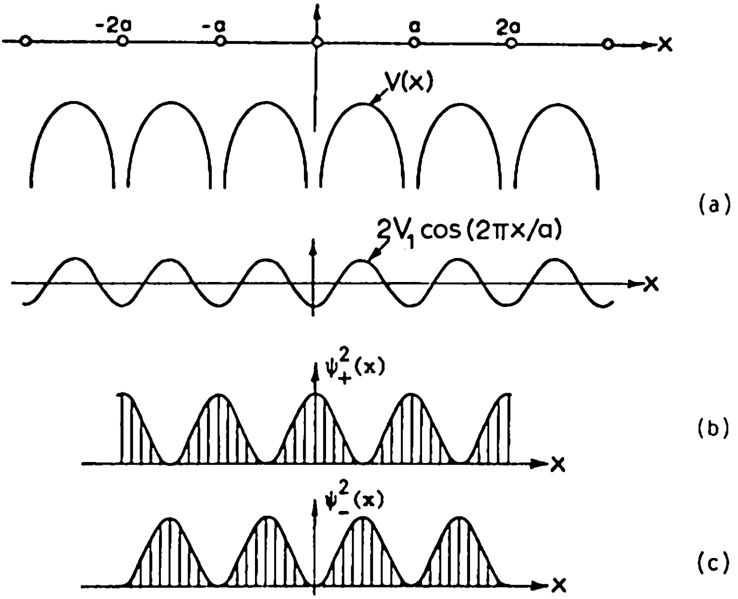


Fig.181.3 (a) The symmetrical periodic potential with the origin at the bond center position between two adjacent host atoms. Note $V_1<0$. (b) The standing wavefunction squared $|\psi_+(x)|^2$ at the lower energy, $E_+=(h^2/2ma^2)+V_1=(h^2/2ma^2)-|V_1|$, of the energy gap (upper edge of the valence band). (c) The standing wavefunction squared $|\psi_-(x)|^2$at the higher energy, $E_-=(h_2/2ma^2)-V_1=(h^2/2ma^2)+|V_1|$, of the energy gap (lower edge of the conduction band).

Let the coordinate origin ($x=0$) be located at a minimum of the periodic potential or coincident with an atomic core as shown in Fig.181.3(a). Then the wavefunction $\psi_+(x)=2C_0\cos(\pi x/a)$ is obviously the solution at the lower total energy E_+ since the electron is concentrated at $x=0$ where $V(x=0)=2V_1\cos(2\pi x/a)=2V_1=-2|V_1|$ from $V_1<0$. Similarly, $\psi_-(x)=i2C_0\sin(\pi x/a)$ has the higher total energy E_-. Because the average kinetic energy of both cases is same as the free electron, $h^2/2ma^2$, the above result is also obtained using only the potential energy.

Another useful property can be obtained from the energy-wave number relationship given by (181.12). This is the dependence of the energy on k near the band edges $k=\pm\pi/a$ where the E-k diagram is a maximum or minimum. The reason of this interest is that the minimum and maximum are the respective conduction and valence band edges of a model semiconductor. Furthermore, near the E-k maximum or minimum, E would depend on k parabolically, i.e. $E \propto k^2$, which is particularly interesting since we recall that for a free electron or plane wave in a constant potential, we have the simple parabolic relationship, $E = \hbar^2k^2/2m$ where m is the free electron mass. Thus, the parabolic E-k relationship near the maximum or minimum E or band edges, suggests that an effective mass can be defined in analogy to the free electron E-k parabola.

Since the first derivative vanishes near a maximum or minimum, the first nonvanishing term in the Taylor series expansion of the E-k relationship near its extrema must be the second or a higher order term. Thus, near $k=K_1/2=\pi/a\equiv k_1$,

$$E(k) = E(k_1)+(dE/dk)(k-k_1)+(d^2E/dk^2)(k-k_1)^2/2!+(d^3E/dk^3)(k-k_1)^3/3!+\dots$$
$$= E(k_1)+ \quad 0 \quad +(d^2E/dk^2)(k-k_1)^2/2!+\dots. \qquad (181.16)$$

The derivatives are evaluated at $k=k_1=K_1/2=\pi/a$. Comparing the second order term of the above expansion with the free electron parabola, $E=\hbar^2k^2/2m$, we see that they are identical if we shift the origin of E-k axes by the amount of $E(k_1)$ and k_1. The E-k relationship in the new coordinate system is just a parabola given by

$$\xi(\kappa) = (d^2\xi/d\kappa^2)^2\kappa^2/2 = \hbar^2\kappa^2/2m^* \qquad (181.17)$$

where

$$\xi = E(k) - E(k_1) \qquad (181.17A)$$

and

$$\kappa = k-k_1. \qquad (181.17B)$$

The effective mass is defined by

$$m^* = \hbar^2/(d^2E/dk^2) = \hbar^2/(d^2\xi/d\kappa^2). \qquad (181.18)$$

The second derivative is evaluated at $k=k_1$ or $\kappa=0$. Using this relationship, the effective mass can be obtained from (181.12). There are two effective masses at $k=k_1$, one at conduction band edge, E_+ or E_C, and the other at the valence band edge, E_- or E_V. Taking d^2/dk^2 of (181.12) and setting $dE/dk=0$ at $k = k_1 = K_1/2 = \pi/a$, with $E_1=\hbar^2k^2/2m$, we then get

$$d^2E/d(\hbar k)^2 = (E+E_1)/(E-E_1)m.$$

Then, the two effective masses are

$$\begin{aligned} & m_+/m = +\ 1/[+1 + 2\hbar^2k_1{}^2/mE_G] \quad \text{at } E=E_+=E_C \\ \text{and} \quad & m_-/m = -\ 1/[-1 + 2\hbar^2k_1{}^2/mE_G] \quad \text{at } E=E_-=E_V. \end{aligned} \tag{181.19}$$

where $E_G=2|V_1|$ is the energy gap. The energy $2\hbar^2k_1{}^2/m = 2\hbar^2(\pi/a)^2/m$ above is slightly more than four times the valence bandwidth, 4BW, since the energy gaps subtract out a small amount of the free-electron parabola as illustrated in Fig.181.2(b). Note $|m_-|>|m_+|$ or hole is heavier than electron.

Usually, the energy gap is much smaller than the allowed bandwidth (in Si, $E_G=1.2$eV and CBW$\propto$4eV and VBW$\propto$12eV) so that the term $2\hbar^2k^2/mE_G$ in the denominator of the two masses (181.19) is much greater than 1. This gives an interesting and expected result: one of the masses, namely m_-, is negative. We would have expected this result earlier and intuitively since the effective mass is inversely proportional to the curvature of the E-k relationship, and the curvature is negative near a E-k maximum. Figure 181.2(b) shows that the lower energy $E_-=E_V$ at $k=\pi/a$ is a maximum and thus has a negative curvature. Thus, we would have expected a negative effective mass for electrons at this energy level or the top edge of the valence band. This result introduces the notion that near the top of the valence band, a negatively charged electron moves as if it has a negative effective mass, or it moves in the same direction as a positively charged particle with a positive effective mass. Such a negative mass and positively charged electron is known as a hole. The concept of hole will be further described in section 190 by calculating the electric current due to a hole.

182 The Tight-Binding Model

The parameters just obtained from the nearly-free electron model are physically related to the properties of the constituent atoms through the Fourier series expansion of the periodic potential. The periodic potential is built up by summing the potential of each atomic core. In principle, the electron-electron repulsive potential of the valence electrons must also be included in the periodic potential which is a very difficult problem. The connection between the resulting energy band and energy gap parameters and the atomic property of the constituent atoms is demonstrated in the preceding section. It is shown there that the energy gap is twice the Fourier series expansion coefficient of the periodic potential and the effective masses near the energy band edges are roughly proportional to the energy gap. Also, $|m_-|>|m_+|$ or hole is heavier than electron.

The tight-binding model gives a still simpler picture and just as quantitative. It provides a direct demonstration of the formation of energy bands from discrete energy levels of isolated atoms. However, the physical picture needed to justify the mathematical approximations used to obtain analytical solutions in the tight-binding model is harder to visualize. We shall first give a simple illustration of how the energy bands are formed. The mathematical details of the tight-binding analysis will then be sketched to complete the discussion.

As we have indicated in Fig.180.1, the tight-binding model is better suited to describe the tightly bound electrons such as the electron inside the deep potential trough labeled (2) in this figure. We shall use the uncertainty principle $\Delta E \Delta t \propto \hbar$ and the electron tunneling model to show that an energy band is formed when atoms are strung together to form a one-dimensional crystal.

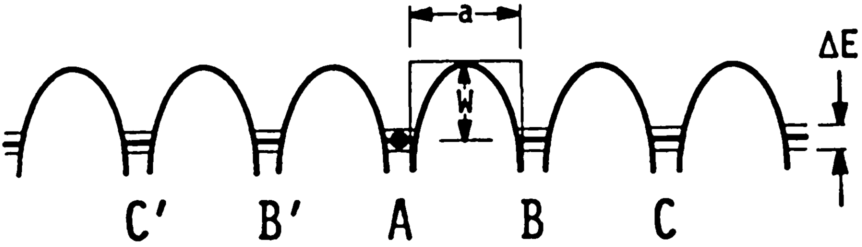


Fig.182.1 Periodic potential of an electron in the tight-binding model to illustrate the formation of energy band from uncertainty principle.

Consider an electron around atom A shown by a dot in Fig.182.1. Its energy does not have a discrete value because the electron at A has a finite probability of showing up at atoms B or B', C or C', and other atoms further away. Hence, the electron is not permanently and completely bound to atom A while it would be around one isolated atom. To estimate the energy uncertainty of this electron, we use the uncertainty principle $\Delta E \Delta t \propto \hbar$ where $(\Delta t)^{-1}$ is the probability per unit time that the electron at A will tunnel through the potential hills into B or B'. Its stay at B is also not definite. Thus, $(\Delta t)^{-1}$ is the probability that it is not at A, B or C, etc. but it is located somewhere in-between the atoms. Suppose that the total wavefunction is given by $\psi = \psi_A + \psi_B + \psi_{B'} + \psi_C + \psi_{C'} + \ldots$ and to be specific, we approximate these by the 1s atomic wavefunctions $\psi_A = \exp(-\alpha x)$, $\psi_B = \exp[-\alpha(a-x)]$ for $0 < x < a$ and $\psi_B = \exp[-\alpha(x-a)]$ for $x > a$, etc. Then, the total probability of finding the electron per unit length at any point in the crystal is

$$P(x) = \psi^*\psi = \psi_A^2 + \psi_B^2 \ldots\ldots + 2\psi_A\psi_B + 2\psi_A\psi_C \cdots .$$

Here, ψ_A^2 is the probability of finding the electron around A atom, $2\psi_A\psi_B$ is the probability of finding the electron neither at A nor at B but between A and B atoms. Thus, $1/\Delta t$ is roughly given by integrating the product $\psi_A\psi_B$ over the entire crystal. The energy uncertainty can then be estimated as follows:

$$\Delta E \propto \hbar/\Delta t = 2\hbar v \int \psi_A \psi_B dx = 3(\hbar v/\alpha a)\exp(-\alpha a)$$

where $\alpha \approx \hbar^{-1}\sqrt{(2mW)}$ is estimated from the square well approximation to the potential hill between atoms A and B shown in Fig.182.1. v is the velocity or frequency with which the electron at A is hitting the wall of the well trying to escape to position B. Although the estimate given above is hand wavering and not rigorous, it gives a general illustration showing that a taller (W) and wider (a) potential well gives smaller energy uncertainty or smaller energy band width, ΔE.

The mathematical analysis used to give an approximate analytical solution in the tight-binding method has its origin in the perturbation theory of differential equations. What is known here is the solutions of the electron wavefunctions in an isolated atom. The effect of the other electrons and nuclei from the adjacent or surrounding atoms is treated as a perturbation to the solutions in an isolated atom. Thus, to start, we assume that the solution of the isolated atoms are known. For example, these are given by the self-consistent solutions of the electron wavefunctions and energies of the Schrödinger equation (SE) listed in Tables 162.1 and 162.2 and the solutions of the atomic hydrogen summarized in Table 156.1 and Figs.156.2(a)-(c). These known solutions are designated by $\chi_n(x)$ and $E_n{}^o$ where n is the effective principle quantum number. They satisfy the atomic SE

$$[-(\hbar^2/2m)(d^2/dx^2) + V_a(x)]\chi_n(x) = E_n{}^o\chi_n(x) \qquad (182.1)$$

$$= [\qquad H_a \qquad]\chi_n(x) \qquad (182.1A)$$

where the atomic potential energy $V_a(r)$ is that of the atom located at the origin, $r=0$, and the sum of the kinetic energy operator and the potential energy is expressed by one symbol, H_a known as the atomic Hamiltonian operator. To implement the perturbation analysis, this atomic SE of one isolated atom (whose solutions are assumed to be known and given) is compared with the crystal SE of many atoms in a semiconductor (whose solutions we are seeking)

$$[-(\hbar^2/2m)(d^2/dx^2) + V(x)]\psi(x) = E\psi(x) \qquad (182.2)$$

$$= [\qquad H \qquad]\psi(x). \qquad (182.2A)$$

V(x) is the periodic potential (energy) from the many atoms and electrons in a crystal. The notation is again simplified by expressing the sum of the kinetic energy operator and the potential energy by the symbol H (the crystal Hamiltonian operator).

We immediately notice that we can rewrite the crystal SE, (182.2), by extracting out the atomic potential from the crystal periodic potential, $V(x) = V_a(x) + [V(x)-V_a(x)] = V_a(x) + V'(x)$. The excess potential, $V'(x)=V(x)-V_a(x)$, comes from the presence of all the other atoms and electrons and is treated as a perturbation to the electron states on the isolated atom located at $r=0$. The crystal SE is manipulated so that the crystal H is split into two terms:

$$\{[-(\hbar^2/2m)(d^2/dx^2) + V_a(x)] + V'(x)\}\psi(x) = E\psi(x) \qquad (182.3)$$

$$= \{[\qquad H_a \qquad] + V'(x)\}\psi(x). \qquad (182.3A)$$

The excess or perturbation potential, V'(x), is illustrated in Figs.182.2(a) and (b). Figure (a) shows the atomic potential (potential energy of an electron in an isolated atom), $V_a(r-R_j)$, of the atom located at $R_j=ja1_x$. Figure (b) shows the excess or perturbation potential, $V'(r-R_j)$.

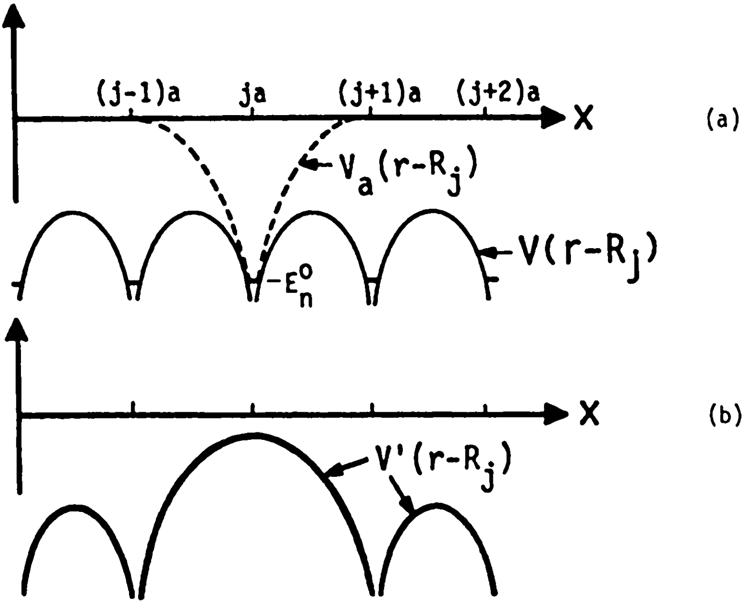


Fig.182.2 (a) The electron potential energy in an isolated atom, $V_a(r-R_j)$, and in a crystal, V(r). (b) The excess or perturbation potential, $V'(r-R_j)$.

The rearranged SE of the crystal, (182.3) and (182.3A), is then solved by expanding the crystal wavefunction, $\psi(x)$, in terms of a linear combination of the known and given atomic wavefunction, $\chi_n(x)$. This is the origin of the term, **linear combination of atomic orbitals** or LCAO. The atomic wavefunctions themselves are inadequate since it does not take into account the phase shift from one atom to the next which is required by the Bloch Theorem. Multiplying the atomic wavefunction by the phase shift factor, $\exp(ikR_j)$, will meet this requirement where $R_j=ja$ is the position of the j-th atom for this one-dimensional illustration. Thus, the LCAO expansion is

$$\psi(x) = \sum_n C_n \chi_n(x-R_j)\exp(ikR_j). \qquad (182.4)$$

The standard procedure of solving a differential equation by the method of expansion in another known function is to compute the expansion coefficient. We have already done this when we did the expansion in Fourier series while developing the plane wave or nearly free electron model. Thus, we repeat the procedure here by multiplying the crystal SE, (182.3A), by the complex conjugate $\chi_1^*(r)\exp(-ikR_{j'})$ and integrate over all space. We then get

$$\begin{aligned} 0 &= \int \chi_1^*(x-R_j)\left[H_a + V'(x) - E\right]\psi(x)dx \\ &= \int \chi_1^*(x-R_j)\left[H_a + V'(x) - E\right]\sum_n C_n\chi_n(x-R_j)\exp(ikR_j)dx. \end{aligned} \qquad (182.5)$$

Let us consider only one band, $n=1$, so $C_1=1$ and $C_n(n\neq 1)=0$, and use the normalization property of the atomic wavefunction when integrated over all space:

$$\int \chi_1^*(x-R_j)\chi_1(x-R_j)dx = 1 \qquad (182.6A)$$

We shall also define two other integrals as follows.

$$\int \chi_1^*(x-R_{j'})\chi_1(x-R_j)dx = \alpha(R_{j'j}) \qquad (182.6B)$$

and

$$\int \chi_1^*(x-R_{j'})V'(x)\chi_1(x-R_j)dx = \beta(R_{j'j}) \qquad (182.6C)$$

The integrations are taken over all space. The integrands of these defined integrals are illustrated in the two lowest diagrams in Fig.182.3 which give an indication of the relative magnitude of these integrals.

By using again the property of the atomic Schrödinger Equation given by (182.1A) for $n=1$

$$H_a\chi_1(x) = E_1{}^o\chi_1(x)$$

and retaining only the nearest neighbor terms, $j'=j\pm 1$ or $R_{j'j}=ja-j'a=(j-j')a=\pm a$ as illustrated in Fig.182.3, then, (182.5) gives

$$E = E_1{}^o + \alpha(a) + 2\beta(a)\cos(ka) + \text{higher order terms}. \qquad (182.7)$$

This simple E-k relationship is plotted in Fig.182.4(a) for an assumed lattice spacing of $a=a_1$. The point $k=0$ is labeled by Γ^+. The two equivalent points $k=+\pi/a_1$ and $k=-\pi/a_1$ are both labeled by H^+. This label convention was introduced by group-theoretical solid-state physicists while studying the symmetry properties of the direct and reciprocal lattices of crystals.

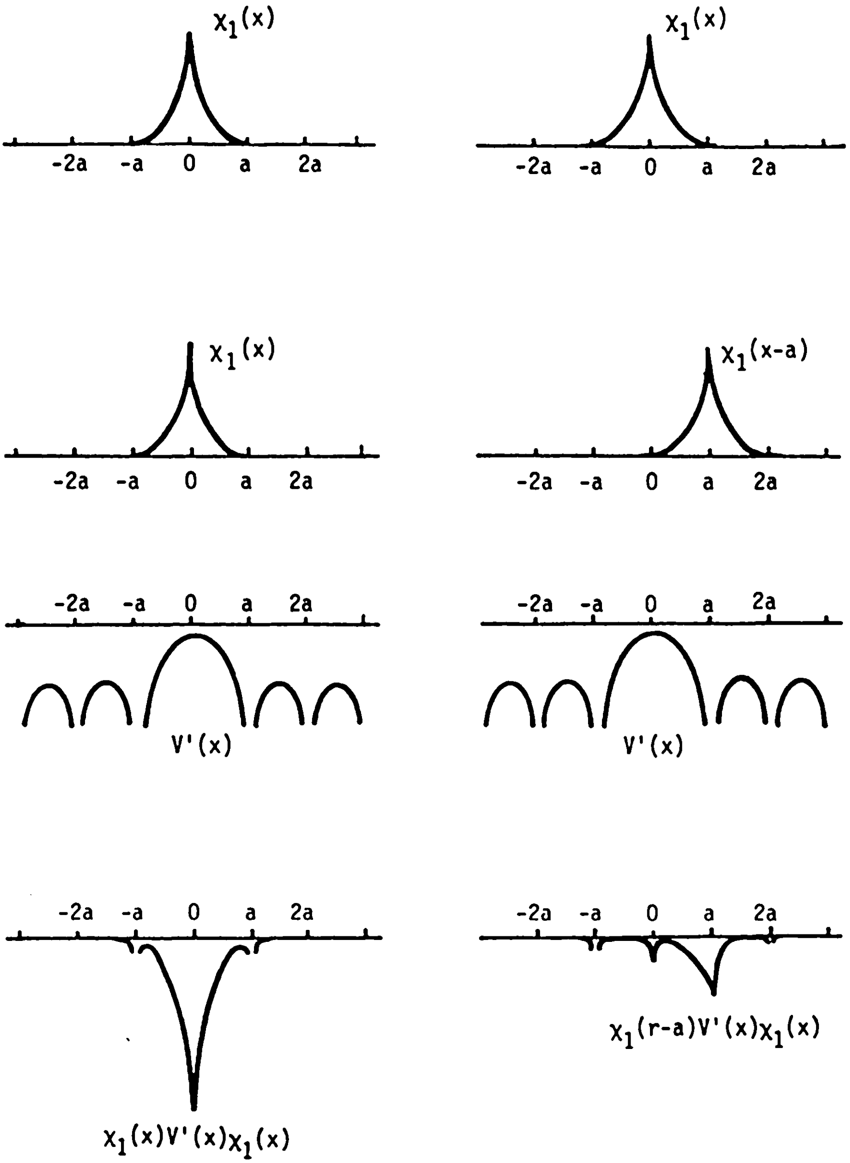

χ1(x)
-2a -a 0 a 2a
χ1(x)
-2a -a 0 a 2a
χ1(x)
-2a -a 0 a 2a
χ1(x-a)
-2a -a 0 a 2a
-2a -a 0 a 2a
V'(x)
-2a -a 0 a 2a
V'(x)
-2a -a 0 a 2a
χ1(x)V'(x)χ1(x)
-2a -a 0 a 2a
χ1(r-a)V'(x)χ1(x)

Fig.182.3 The 1s-like atomic wavefunctions, $\chi_1(x)$ and $\chi_1(x\text{-}a)$, the excess or perturbation potential, $V'(x)$, and the integrands $\chi_1(x)V'(x)\chi_1(x)$ and $\chi_1(x\text{-}a)V'(x)\chi_1(x)$.

The E-k diagram of Fig.182.4(a) shows several important features. (1) The E-k curve is periodic in k with a periodicity $ka_1=2\pi$. (2) The energy levels are shifted from their position in an isolated atom, $E_1{}^o$, by the integral $\alpha(a_1)$ but the energy band clearly arises from the atomic energy level. (3) The allowed bandwidth is given by $4\beta(a_1)$ which represents the effect of the neighboring atoms on the atomic energy level through the perturbation potential V'(x). (4) The effective masses computed from $\hbar^2/(d^2E/dk^2)$ at the band edges or energy extrema ($ka_1=0,\pm\pi$) is given by $\hbar^2/[2\beta(a)a^2]$. Thus, it is inversely proportional to the bandwidth.

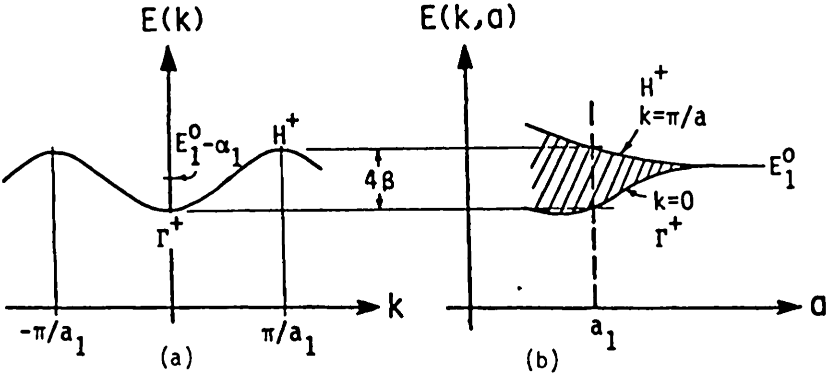


Fig.182.4 The tight-binding electron energy band in a 1-d crystal. (a) The E-k (energy-wave number) diagram from the atomic s-state (n=1) with $\beta>0$ and $a=a_1$. (b) Variation of the band edges at $k=\pm\pi/a_1$ and bandwidth with lattice constant.

The tight-binding method can also demonstrate the shifting of the energy levels with interatomic spacing, a. Figure 182.4(b) shows the variation with 'a' of the band edges H^+ at $k=\pm\pi/a$ and Γ^+ at $k=0$ and the bandwidth (shaded). This figure shows that the tight-binding method can also be used to demonstrate the crossing of the energy levels as the interatomic spacing decreases, such as those shown in Fig.173.1(c). In order to illustrate the crossing of energy levels, obviously two atomic states must be considered instead of just one given in Fig.182.4(b). The results of a two-atomic-level or two-band tight-binding calculation are given in Figs.182.5(a) and (b). The E-a diagram in figure (b) show the two energy bands from two atomic states, n=1 and n=2. The two energy bands, the 1s and 2s bands, are shaded in black to reflect the very large number of energy levels, one from each of the 10^{23} atoms in a 1 cm^3 crystal. The two bands cross at a smaller interatomic spacing than shown in the figure. Note that in this two-band example, the valence band maximum [labeled as E_V previously such in Fig.181.2(a)], H^+, and the conduction band minimum (labeled E_C previously), H^-, are located at $k=\pm\pi/a$. The other band edges, the valence band minimum (labeled as E_V, previously), Γ^+, and the conduction band maximum (labeled as E_C, previously), Γ^-, are located at $k=0$.

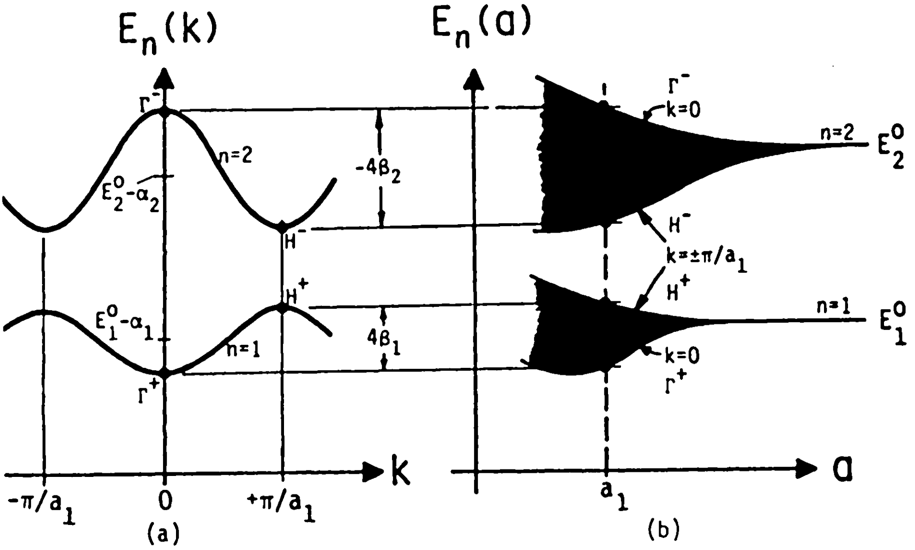


Fig.182.5 The interaction of two energy bands from two atomic states, n = 1 and n = 2. (a) The E-k diagrams. (b) The E versus interatomic spacing a for the band edges, k = 0 and $k=\pm\pi/a$.

One principal and fundamental feature from the tight-binding model is that the location of the energy levels are given using the vacuum level as the reference. This is a very important result since many experiments measure the electron energy relative to the vacuum level. In contrast, the plane wave or nearly-free electron model described in section 181 does not give the absolute energy reference since the average of the periodic potential, V_0 in (181.6), cannot be computed accurately from Fourier expansion because V(x) is not accurately known. Thus, the tight-binding result is more fundamental than the nearly-free electron result. The latter is frequently used as an empirical model that can be used to interpolate the few E-k experimental data points to give the complete E-k diagram over all the k values (from $-\pi/a$ to $+\pi/a$ in the 1-d example). The Fourier expansion parameters of the latter, V_0, V_1, V_2, ... , C_0, C_1, C_{-1}, ... are adjusted so the theoretical E-k fits the few experimental E-k data points.

In practical theoretical computations of the energy bands using only the Coulomb and relativistic forces, the crystal structure and the lattice constant, neither the nearly-free electron model nor the tight-binding model have given numerically accurate energy bands. Accuracy better than 0.1 eV (0.03eV claimed in 1990) has not been achieved even when many (hundreds) plane waves or many (hundred) atoms are used in the largest supercomputer. Computer speed and memory size are the limit which require that the nonlocal potential from the atomic nuclei and the core electrons be approximated by some effective local potential which is no longer fundamental and whose parameters are empirical. The empirical parameters cannot accurately represent the real potential arising from the many nonlocal or spatially distributed core electrons around each atomic core and the many valence electrons distributed over the entire crystal.

Nevertheless, both the nearly-free electron and the tight-binding models are exceedingly useful pedagogical tools to help the students visualize the concept of electrons and holes and **analyze** and **predict** the electrical characteristics of transistors and integrated circuits, provided one is willing to use a few experimental parameters in the model, such as the mobilities, the lifetimes, the effective masses, the energy gap, and others. These are the very parameters whose origin and whose justification for use in device analysis will occupy the next two chapters.

183 Energy Band Diagrams of Semiconductors

The foregoing description of the energy band diagrams using the nearly-free electron and the tight-binding models shows that the energy bands not only are a function of position, x, but also a function of the wave number of the valence electrons in the crystal. These are known as the E-k diagrams while the energy band diagrams in position space described in section 172 are known as the E-x diagrams.

A graphical summary of these two descriptions of the energy bands in Si, E-x and E-k, are given in Figs.183.1(a) and (b) respectively. The E-k diagram in Fig.182.1(b) is based on the theoretical result of fitting a 22 plane-wave three-dimensional nearly-free electron model to the few experimental data of energy gaps and effective masses. The figure gives only the E-k dependences in two k-vector directions, <111> and <100> while the calculation gives the E-k dependences in all k-vector directions. Energies at high symmetry points are labeled by Greek and English alphabets with numeral subscripts, such as X_1 and X_4 at $(k_x,k_y,k_z)=(2\pi/a)(1,0,0)$, and $\Gamma_{2'}$, Γ_{15} and $\Gamma_{25'}$ at $k=0$ which is the convention introduced and used by group-theoretical solid-state physicists.

It is evident from figure (b) that there are many branches in the conduction and valence bands. It illustrates the complexity of the E-k diagram in a real three-dimensional solid. In addition, E_C is located at about $0.85(2\pi/a)(1,0,0)$ or its five equivalent positions in k-space and it is not lined up with E_V which is located at $k=0$ and labeled $\Gamma_{25'}$. This is known as **indirect energy gap** because E_C and E_V are not lined up vertically in k-space or have different k-values.

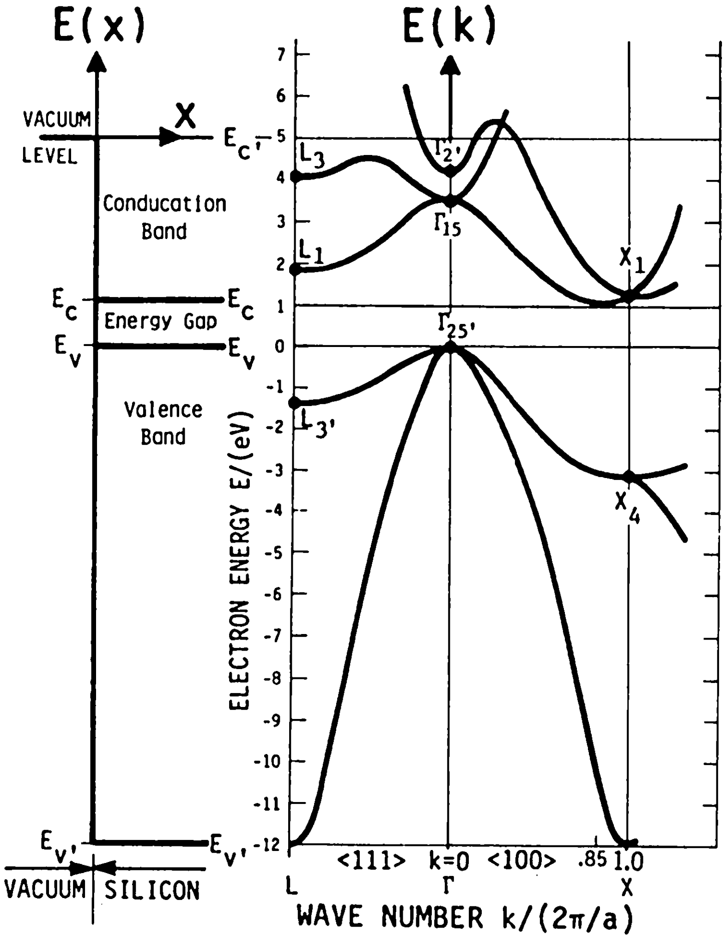

E(x)
VACUUM LEVEL
X
Conducation Band
E_C
Energy Gap
E_V
Valence Band
$E_{V'}$
VACUUM
SILICON
$E_{C'}$
E(k)
ELECTRON ENERGY E/(eV)
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
L_3
L_1
$L_{3'}$
$\Gamma_{2'}$
Γ_{15}
$\Gamma_{25'}$
X_1
X_4
<111> k=0 <100> .85 1.0
L
Γ
X
WAVE NUMBER k/(2π/a)
(a)
(b)

Fig.183.1 The energy band diagrams of silicon crystal. (a) The E-x diagram and (b) the E-k diagram. See text for explanation of the labels.

The relative position of the two band edges, E_C and E_V, in k-space have a dominant effect on the optical properties of the solid. For example, Si is a very poor light emitter because its energy gap is indirect or its electrons are located at a k point very different from that of the holes which is at k=0. Thus, when an electron recombines with or drops into a hole in Si, the large momentum change, $\Delta P=\hbar\Delta k=\hbar(k_e-k_h)=0.85\hbar(2\pi/a)$, cannot be carried away by the photons since photon momentum is too small (1000 times too small). Similarly, a photon alone cannot generate an electron-hole pair efficiently in Si even when the photon has sufficient energy ($>E_G$) because the photon does not have enough momentum to account for the necessary momentum change when an electron at the Si valence band edge is moved up to the Si conduction band edge. A phonon, which is a quantized lattice vibration wave, is needed to conserve the momentum during the optical transitions in an indirect energy gap semiconductor.

Some of the salient features of the E-k diagram of real 3-d crystals, such as Si shown in Fig.183.1, cannot be predicted by the 1-d theory. However, these features in real 3-d crystals are easily predicted by either the 3-d plane-wave (Fourier expansion) or 3-d tight-binding model. For example, in the 1-d theory the extrema of the E-k diagram must be located at $k=0$ or $k=\pm\pi/a$ but in real 3-d (or 2-d) crystals, these extrema can be located at any k point between $k_x=-\pi/a$ and $+\pi/a$, $k_y=-\pi/a$ and $+\pi/a$, and $k_z=-\pi/a$ and $+\pi/a$.

Another feature is the overlapping 3-d energy bands and gaps along different directions of k. For example, the experimental energy gap at L between $L_{3'}$ and L_1 is about 3.4eV but it is overlapped by two branches of allowed bands: the conduction band branch between $\Gamma_{15}(k=0)$ and X_1, and the valence band branch between $L_{3'}$ and $\Gamma_{25'}(k=0)$. Thus, the remaining true energy gap is only 1.2eV. The 1-d theory does not allow any overlap: the adjacent allowed energy bands are always separated by an energy gap. This 1-d feature was the key that motivated the solid state theorists to work out the mathematics of the 2-d and 3-d energy band theory of metals in the 1930's after quantum mechanics was invented. Because 1-d theory, as we shall see in the next section on the energy band of metals, cannot predict the tremendous number of conduction electrons in some conductors as indicated in Table 111.1. For example, elements in group IIA of the periodic table (Be, Mg, Ca, Sr, Ba) have essentially metallic conductivity, $(0.02$ to $0.2)\times10^6$ S/cm or 50 to 5 $\mu\Omega$-cm, with about 10^{21}-10^{22} electron/cm^3. But the 1-d energy band theory predicts zero electron in the upper or conduction band and zero hole or completely filled lower or valence band when the host atom has two valence electrons, such as the two ns^2 electrons in group IIA elements. As illustrated by the Si example in Fig.183.1(b), the 3-d theory allows the overlap of energy bands for energy extrema at different k points. This removes the necessary existence of an energy gap between two adjacent extrema or adjacent maximum and minimum which is demanded by the 1-d theory. Thus, 3-d theory is mandatory for predicting metallic conduction and 1-d is inadequate.

Figure 183.2 gives the historical first theoretical E-k diagrams of many elemental and binary compound semiconductors. They were systematically computed by M.L.Cohen and T.K.Bergstresser of the University of Chicago in 1965 and published in January 1966.

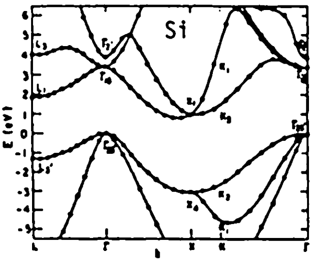


Fig. 1. Band structure of Si.

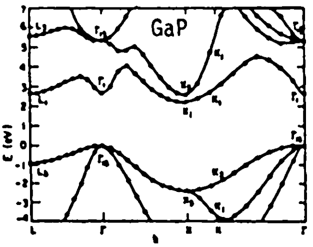


Fig. 4. Band structure of GaP.

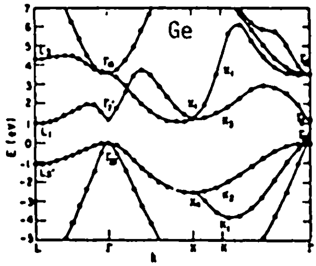


Fig. 2. Band structure of Ge.

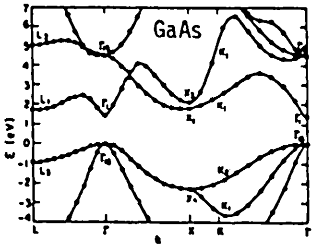


Fig. 5. Band structure of GaAs.

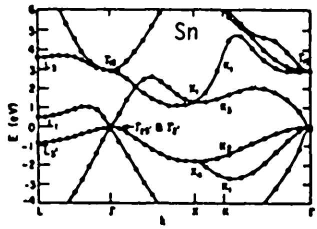


Fig. 3. Band structure of Sn.

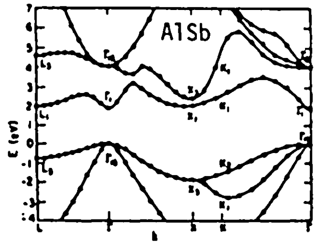


Fig. 6. Band structure of AlSb.

Fig.183.2 The computed 'pseudo-potential' or empirical potential energy bands of fourteen semiconductors by Cohen and Bergstresser using Fourier expansion. (Continued on next page.)

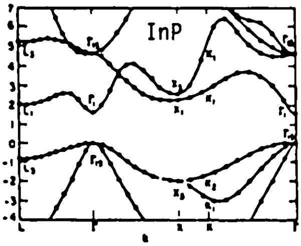


Fig. 7. Band structure of InP.

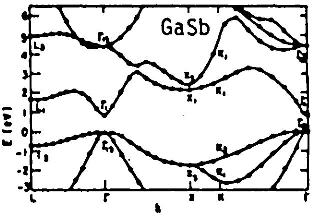


Fig. 8. Band structure of GaSb.

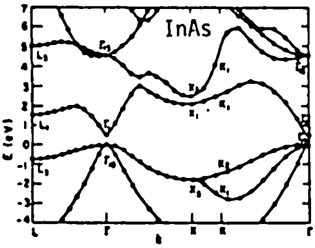


Fig. 9. Band structure of InAs.

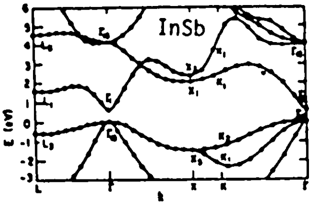


Fig. 10. Band structure of InSb.

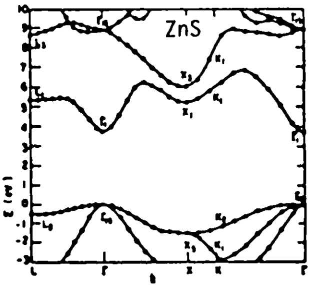


Fig. 11. Band structure of ZnS.

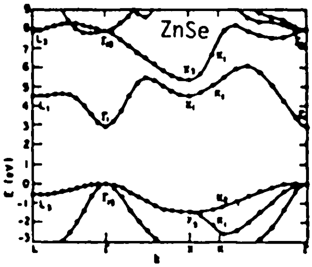


Fig. 12. Band structure of ZnSe.

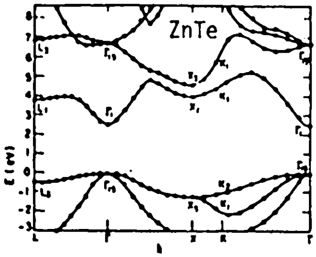


Fig. 13. Band structure of ZnTe.

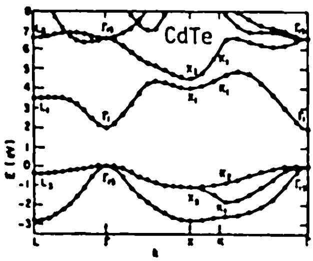


Fig. 14. Band structure of CdTe.

Fig.183.2 The computed 'pseudo-potential' or empirical potential energy bands of fourteen semiconductors by Cohen and Bergstresser using Fourier expansion. (Following previous page.)

Cohen and Bergstresser used the 3-d nearly-free electron model and the Fourier series analysis. The input data are experimentally measured vertical energy gaps at different k mostly from optical absorption experiments. The numerical values of the lowest three Fourier expansion coefficients of the periodic potential were empirically obtained by least-squares fitting the experimental energy gaps to a 22x22 matrix from using 22 plane waves or first 22 terms in the infinite Fourier series. Our 2-plane wave example in section 181 had a 2x2 matrix. Thus, the calculated results are empirically fitted energy bands instead the true or first-principle theory which would have used Coulomb's law, relativistic mechanics (spin-orbit force and relativistic mass), the crystal structure, and no experimental energy band data. Nevertheless, the empirical least-squares fitted results of Cohen and Bergstresser do predict the trend of many optical and electrical properties of these semiconductors. They also help to anticipate and even design new materials such as impurity-doped semiconductors and multi-element (binary, tertiary,...) compound semiconductors, ionic insulators, and metals.

Let us briefly discuss some of the results shown in Fig.183.2. The E-k diagram of GaAs shows that its CB minimum and VB maximum are both located at k=0. This is known as the **direct energy gap**. It is the fundamental reason that GaAs is an excellent light emitter, giving highly efficient light emitter diodes and diode junction lasers because photon momentum is very small.

These early computations also showed that GaP has an indirect energy gap while ZnS, ZnTe and CdTe have direct energy gaps.

Another systematic and general trend confirmed by these empirical calculations is that the more ionic the compound, reflected by the larger core difference between the two host atoms, the larger the energy gap. For example, consider the isocoric (coined by S.T.Pantelides and this author in 1970) and isoelectronic series, Ge (IV), GaAs (III-V), ZnSe (II-VI) and CuBr (I-VII), the energy gap increases as Ge(0.67eV), GaAs(1.35eV), ZnSe(2.6eV) and CuBr(>3eV).

This systematic trend of increasing energy gap is also evident in the covalent series, C(5.4eV, not computed by C&B but computed by later researchers), Si(1.2eV), Ge(0.7eV), and α-Sn(-0.08eV). This trend answers a fundamental question. The valence electron model is inadequate to explain the differences in energy gaps and in the detailed electrical, mechanical and optical properties. **It is the charge distribution of the atomic core (nucleus plus the core electrons) that accounts for these differences.** Coulomb force is still the fundamental and dominant glue. Relativistic effects (spin-orbit force and relativistic mass) will give additional but not very large modifications of the properties.

184 Energy Band of Metals and Conductors

The high conductivity property of metals cannot be accounted for by the 1-d energy band model. In the preceding section, we have briefly described the most serious contradiction which is now discussed in more detail. The 1-d energy band model has the unique property that two adjacent allowed energy bands must be separated by a forbidden energy gap. From section 173 on filling of the energy levels in the allowed bands by the available valence electrons, we immediately arrive at the conclusion on the conductivity magnitude listed in Table 184.1 for the crystals with one to six valence electrons based on the 1-d model.

TABLE 184.1
Band Filling and Conductivity of Solids
Based on the 1-D Energy Band Model

VALENCE ELECTRON OF HOST	NUMBER OF STATES per VALENCE ELECTRON in ONE ALLOWED BAND	FRACTION OF THE ALLOWED BANDS FILLED			CONDUCTOR or INSULATOR
		Lowest	Higher	Next Higher	
*********	*********************	******	******	*********	***********
1	2	0.5	0.0	0.0	CONDUCTOR
2	2	1.0	0.0	0.0	INSULATOR
3	2	1.0	0.5	0.0	CONDUCTOR
4	2	1.0	1.0	0.0	INSULATOR
5	2	1.0	1.0	0.5	CONDUCTOR
6	2	1.0	1.0	1.0	INSULATOR

The four columns on the right hand side of the above table show that the 1-d energy bands are either half-filled when there is an odd number of valence electrons per host atom or completely filled when there is an even number of valence electrons per host atom. Thus, 1-d solids from elements with even number of valence electrons are all insulators because the lower valence bands are filled and the upper valence bands (known as the conduction bands in the semiconductor terminology we have used in the preceding sections) are all empty. Thus, there are neither electrons in the conduction band nor holes in the valence band to give conduction at low temperatures. This 1-d theoretical prediction contradicts the experimental observations which indicate that solids from the group IIA (Be, Mg, Ca), IIB (Zn, Cd, Hg), IVA (Ti, Zr), VIA (Cr, Mo, W), VIB (S, Se, Te), and VIIIA (Re, Co, Ni, Pd, Pt) elements in the periodic table are good conductors and exhibit metallic behavior, i.e. their high room-temperature conductivity persists at low temperatures, and the number of conduction electrons is large and essentially constant, independent of temperature.

The preceding description of the discrepancies between the 1-d energy band theory and experiments led the solid state physicists to begin computing the 3-d energy band structures in the 1930's using realistic crystal structures known at the time. Analytical approximations had to be used to simplify the numerical calculation because high-speed electronic computers had not been invented. The first of the many energy band calculations was made in 1933 by Wigner and Seitz during Seitz's doctoral research at Princeton. They treated it as a boundary value problem. To define the geometrical boundary, they picked the metallic sodium because among metals it has the simplest crystal structure geometry (BCC, body-center cubic) that gives the simplest boundary shape to ease the algebra arising from requiring that the assumed solution satisfies the boundary conditions. In addition, it has only one atom per primitive unit cell which further simplifies the mathematics. However the primitive parallelepiped unit cell shown in Fig.184.1(a) is not a simple geometry for mathematical approximation. But the nonprimitive body-centered cubic unit cell, shown in Fig.132.1(13), is geometrically simpler and rather symmetrical. It has one host atom located at the center of the cube and eight host atoms at each corner of the cube or two atoms per cell.

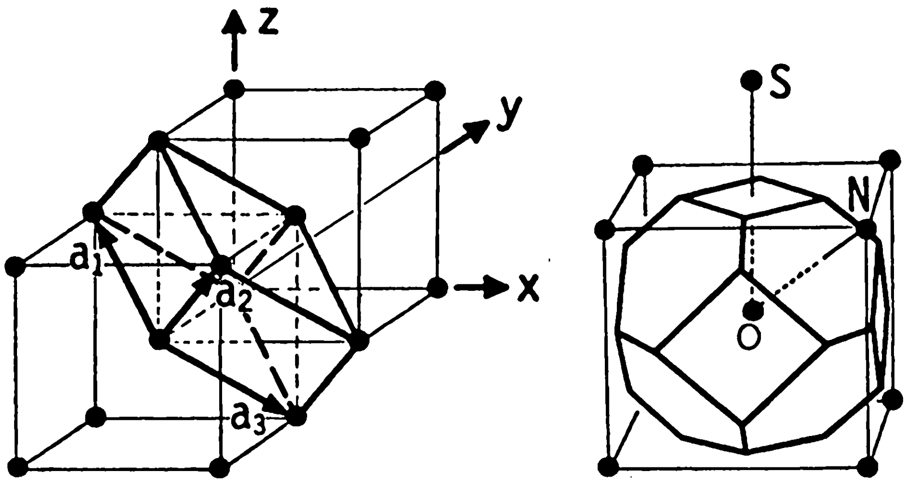


Fig.184.1 The primitive direct unit cells of the body-center cubic lattice, which are identical to reciprocal unit cells of the face-centered cubic lattice. (a) The parallelpiped primitive unit cell. (b) The Wigner-Seitz primitive unit cell. [The cubic unit cell was given in Fig.132.1(13)].

This simple cubic unit cell led Seitz and Wigner to invent a mathematically simple and yet accurate analytical approximation to the atomic core potential by constructing a geometrically simpler primitive unit cell shown in Fig.184.1(b) known as the **Wigner-Seitz** (W-S) cell. Note that it has fourteen surfaces: six square surfaces and eight hexagonal surfaces. The square surfaces are from the six perpendicular bisector planes of the center atom, labeled O, and the six second nearest neighbor atoms, one of which is labeled S. The hexagonal surfaces are from the eight perpendicular bisector planes of the center atom, labeled O, and the eight nearest neighbor atoms on the eight corners of the cube, one of which is labeled N. This 14-face W-S cell is also known as the **truncated octahedron**. The truncation of the 8-equilateral-triangular-faced octahedron by the six perpendicular bisector planes of the lines O-S gives rise to the additional six planes.

Wigner and Seitz noted that since the 1/r Coulomb potential has spherical symmetry, the orthogonal functions known as the spherical harmonics can be used as the basis function to expand the electron wavefunctions around each atomic core. To use this spherical symmetry property to simplify solution matching at the boundaries, Wigner and Seitz first constructed the primitive unit cell shown in Fig.184.1(b), the W-S cell, because it has nearly spherical symmetry. They then divided this W-S cell into two regions by means of the surface of a sphere inscribed inside the W-S cell (not shown). Finally, they required that the solution, a series of spherical harmonics centered at O, satisfies the periodicity condition (repeating itself from one W-S to the next) at the contact points between the adjacent spheres, i.e. the intersection of the O-N line and its perpendicular bisector plane or the hexagonal face shown in Fig.184.1(b). In the exact solution, one would have to satisfy the periodicity condition on all fourteen surfaces of the W-S cell. This method was coined by Wigner and Seitz as the **cellular method** as suggested by the geometrical feature of this boundary value problem. An extension of the cellular method was made by John C. Slater and his graduate students at M.I.T. and extensive energy band calculations began in the 1950's and ended in the 1970's when Slater died after retirement at the University of Florida. Slater coined his extension the Augmented Plane Wave (APW) method. APW makes several extensions to the original W-S cellular method: (i) the solution in the space exterior to the W-S sphere but inside the W-S cell is expanded by plane wave or Fourier series, (ii) the W-S radius is made smaller than the inscribed sphere so the exterior (to the W-S sphere) plane wave solution is matched to the interior spherical harmonics solution over the entire spherical surface of the W-S sphere, and (iii) the radius of the W-S sphere is made an adjustable parameter to fit theory to experimental data.

For a comprehensive review of the early history of the theory and experiments on energy bands in solids, see the 1987 Dover re-publication soft-cover edition of the 1940 classic written by Frederick Seitz, **The Modern Theory of Solids**, Dover Publications, New York (1987). For the early as well as later theoretical developments on energy bands in atoms, molecules and solids, contributed by J.C. Slater and later authors, see his six-volume classic, **Quantum Theory of Matter** (1968), **Quantum Theory of Molecules and Solids, Volume I** (1963), **Volume II** (1965), **Volume III: Insulators, Semiconductors, and Metals** (1967), and **Quantum Theory of Atomic Structure, Volumes I** and **II**(1960), McGraw-Hill Book Company, New York.

It is evident that the two cellular methods just described are tight-binding or extended tight-binding methods with an unique but arbitrary or empirical feature, that of dividing the crystal into two groups of regions: the atomic cores (the cell and hence the term cellular) and the exterior to the cores. This division gives rise to the requirement of matching the solutions on the surface separating the two groups of regions. This matching of two series determines the magnitude and shape of the numerical E-k solutions. The regular tight-binding method and the 3-d plane wave method are mathematically simpler because they do not artificially divide the crystal into two groups of regions. They write one series for the whole crystal, thus, there are no boundary conditions to satisfy on the imaginary interior boundaries.

We shall describe the result of the face-centered cubic (FCC) lattice in detail in the following paragraphs using the plane-wave method. The reason for this choice is that the basic crystal structure of the two elements which dominate the silicon microchip technology, Si semiconductor and Al metal, have the FCC symmetry. Aluminum metal crystallizes in the FCC lattice. Silicon semiconductor crystallizes in the diamond lattice which is composed of two FCC's displayed by a/4 in the [111] direction which was illustrated in Fig.132.2(b).

The algebra from Fourier expansion of a 3-d periodic potential and 3-d wavefunction in the 3-d FCC lattice is straightforward. In fact, it is identical to the 1-d illustration already given in section 181; only the algebra is slightly longer, requiring one to keep track of the three Cartesian components of the wave vector **k** and translation vector of the direct and reciprocal lattices, $\mathbf{R}_n$ and $\mathbf{K}_h$.

In analogy with the 1-d E-k diagram given by the free-electron parabola shown in Figs.181.1(b) and 181.2(b), the 3-d E-k free-electron parabolas are first drawn in several selected directions in the 3-d k-space. Then, energy gaps are opened and E-k distortions are introduced in these continuous E-k parabolas by the periodic potential. The equation of these 3-d E-k parabolas is

$$E(\mathbf{k}) = (\hbar^2/2m)[(k_x+K_{hx})^2 + (k_y+K_{hy})^2 + (k_z+K_{hz})^2]. \tag{184.1}$$

The reciprocal lattice vector, $\mathbf{K}_h$, in (184.1) can be obtained from the given direct lattice translation vector, $\mathbf{R}_n$, of the FCC lattice using

$$\mathbf{K}_h \cdot \mathbf{R}_n = 2\pi x(\text{integer}). \tag{184.2}$$

Using (184.2) one can readily show that $\mathbf{K}_h$ of the FCC lattice is identical to the $\mathbf{R}_n$ of the BCC lattice. Thus, $\mathbf{K}_h$ can be written down by visual examination of the BCC unit cell shown in Fig.184.1(a), which is

$$\mathbf{K}_h = (2\pi/a)[(-h_x+h_y+h_z)\mathbf{i}_x + (h_x-h_y+h_z)\mathbf{i}_y + (h_x+h_y-h_z)\mathbf{i}_z]. \tag{184.3}$$

The h's are positive or negative integers including zero.

From the E-k equation (184.1) and using (184.3), the E-k parabolas of the FCC lattice can be calculated with a slide rule and are shown in Fig.184.2(d). The energy or y-axis is normalized where $\xi_F = (k_F a/2\pi)^2$. The directions in the reciprocal lattice or k-space are shown by the (k_x, k_y, k_z) Cartesian coordinate in Fig.184.2(c). The symbols are those used by group-theoretical solid-state physicists to simplify the discussion and communication on the points and directions in the 3-d k-space. The truncated octahedron in Figs.184.2(a) and (c) is known as the **first Brillouin Zone**. It is the 3-d analogy of the restricted k-space, $-\pi/a < k < +\pi/a$, in the 1-d E-k diagram, such as Figs.182.4(a) and 182.5(a) or the half-size E-k diagram of Fig.181.2(b). Notice that the geometry of the first Brillouin zone of the FCC lattice (the truncated octahedron) is identical to the geometry of the Wigner-Seitz cell of the BCC lattice indicated in Fig.184.1(b). The second 3-d Brillouin zone, in analogy to the spaces between $ka = \pi/2$ to π and $-\pi/2$ to $-\pi$ of the 1-d E-k diagram, has a more complex geometry with $6 \times 4 + 8 \times 6 = 72$ surfaces as shown in Fig.184.2(b).

Two different combinations of h_x, h_y and h_z that give the same numerical sum in the brackets, ( ), of (184.3) would give the same K_h and hence the same energy from (184.1). The two different combinations would give two plane waves of different space orientations or configurations as indicated by the plane wave function

$$\psi(r) = A \cdot \exp[i(k + K_h)] \qquad (184.4)$$

or its linear combinations that give sine and cosine functions. Thus, such a E-k branch would have a two-fold configuration degeneracy. The configuration degeneracy are given in the bracketed numeral next to each of the lower free-electron parabolas in Fig.184.2(d).

These free electron energy-momentum parabolas are the wave number dependences of the electron energy when the core potential of host atoms is hypothetically diminished to zero. There is still a host or ghost lattice. This free electron parabola methodology was first introduced by William Shockley in his Ph.D. thesis under Slater at M.I.T. in the mid-1930s. [See W. Shockley, Physical Review 52, 866 (1937).] Shockley introduced this to test the Wigner-Seitz cellular method by a known solution, the free-electron parabolas. It has been used ever since as a most powerful illustration to visualize the effect of the periodic potential on the free-electron energy. It was coined by Shockley as the **empty lattice bands** since the atomic potentials are diminished to zero or the lattice is empty and not occupied by atoms although the lattice is there. The resultant effect from the presence of a periodic potential is similar to that illustrated for the 1-d case shown in Fig.181.2(b): an energy gap was opened up at $k = +\pi/a$. There is also an energy gap symmetrically located at $-\pi/a$ but it was not shown in the introductory Fig.181.2(b).

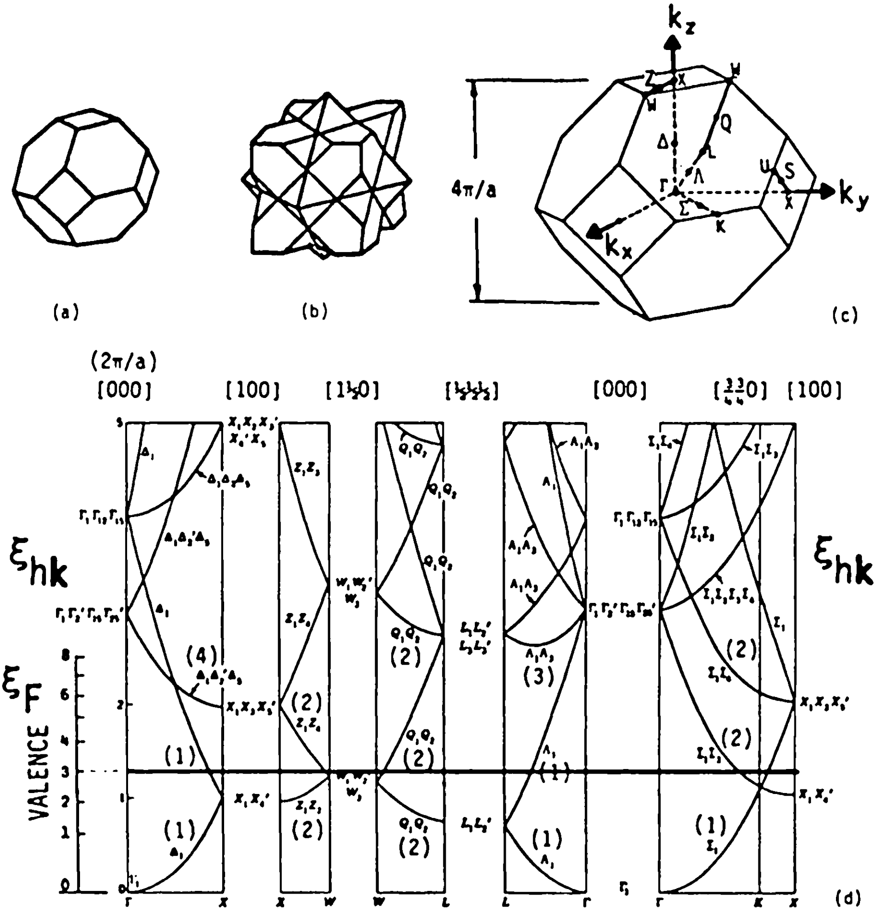

(a)
(b)
(c)
k_z
k_y
k_x
4π/a
Γ
X
W
L
K
U
Δ
Λ
Σ
Q
S
Z
(2π/a)
[000]
[100]
[1½0]
[½½½]
[000]
[¾¾0]
[100]
ξ_{hk}
ξ_{hk}
ξ_F
VALENCE
0
1
2
3
4
6
8
(1)
(2)
(3)
(4)
Γ
X
X
W
W
L
L
Γ
Γ
K
X
(d)

Fig.184.2 FCC lattice. (a) The first and (b) second Brillouin zones. (c) The k point and direction symbols. (d) **The empty lattice bands**. Normalizations are: $\xi = E/[(\hbar^2/2m)(2\pi/a)^2]$, $\kappa_x = k_x a/2\pi$, and $\xi_F = (k_F a/2\pi)^2 = (12V/8\pi)^{2/3}$ which is the normalized Fermi energy.

To illustrate the effect of the atomic potential from the host atoms in a crystal on the free-electron parabolas, we take the case of Al which has three valence electrons. To determine the location of the Fermi level (the energy below which all levels are occupied by electrons) in the various k directions, we draw a Fermi sphere with radius k_F which is determined by $E_F = \hbar^2 k_F{}^2/2m$ and fill up the sphere with the available valence electrons. Knowing the unit cell size, a, and hence the atomic density of Al, the valence electron density can be computed. The Fermi energy can then be computed. The solution is given by $\xi_F = (k_F a/2\pi)^2 = (12V/8\pi)_2{}^{/3}$ where V is the valency of the host atom. Fermi energy level for valence electrons from V=1 to 8 are labeled along the vertical or energy axis in Fig.184.2(d). For aluminum, V=3 and the numerical value of E_F is 11.65 eV for a valence electron concentration of 1.81×10^{23} cm^{-3} (see also section 413 for less normalized computation).

The horizontal ξ_F line at V=3 in Fig.184.2(d) for Al cuts through several E-k parabolas. It shows that the lowest energy bands (first Brillouin zone) are completely filled. The second lowest energy bands (second Brillouin zone) are not quite completely filled, and the unoccupied states in the upper part of the energy bands in this zone give conduction holes. The third Brillouin zone is only partially filled and it accounts for most of the conduction electrons in aluminum. A very small number of electrons at the symmetry points W are spilled over into the fourth Brillouin Zone.

The foregoing 2-d illustration on filling the states and zones by valence electrons can be described more visually using the 3-d E-k energy surfaces first constructed by Walter Harrison in 1960, known as the **Fermi Surfaces of Metals.** These surfaces for the four lowest Brillouin zones in aluminum are given in Fig.184.3. The surfaces in the higher zones are folded back into the first or reduced Brillouin zone. For example, the extended topology of the second Brillouin zone is shown in the lower part of the figure, whose surface planes are partially cut by the Fermi sphere which are not shown in this picture but are folded back into the first or reduced zone and shown the picture above it. The topology of the Fermi surfaces was an exciting and appealing graphical subject during the 1960's when the Fermi surfaces of metals were investigated experimentally by cyclotron resonance experiments. Regions with convex surfaces are occupied by electrons while regions with concave surfaces are empty, not occupied by electrons or occupied by holes. These theoretical shapes were used to analyze the experimental data. The 3-d picture of the four Brillouin zones given in Fig.184.3 again shows the description just given using the 2-d E-k diagram. That is, the first zone is completely filled, the second zone contributes to holes, the third zone is responsible for the majority of the conduction electrons, and the fourth zone contributes a very small amount of conduction electrons. The significance of completely and partially filled energy bands or Brillouin zones is discussed and illustrated in the next section which shows that a completely filled band cannot carry a current and hence give no cyclotron signal or electrical conduction.

In the above description for Al, we have discussed only the consequences of using the free electron parabolas or the empty lattice bands. In a real crystal, the atomic potential will distort and shift the free electron bands and open energy gaps. These changes will modify the band filling picture of Al just described. For example, it can cause the energy bands in the second Brillouin zone to be completely filled, leaving no holes; the bands in the third Brillouin zone to be filled with slightly fewer electrons; and no electrons spilled over to the bands in the fourth Brillouin zone if the energy at the W points are shifted up by the periodic potential from the Al atoms.

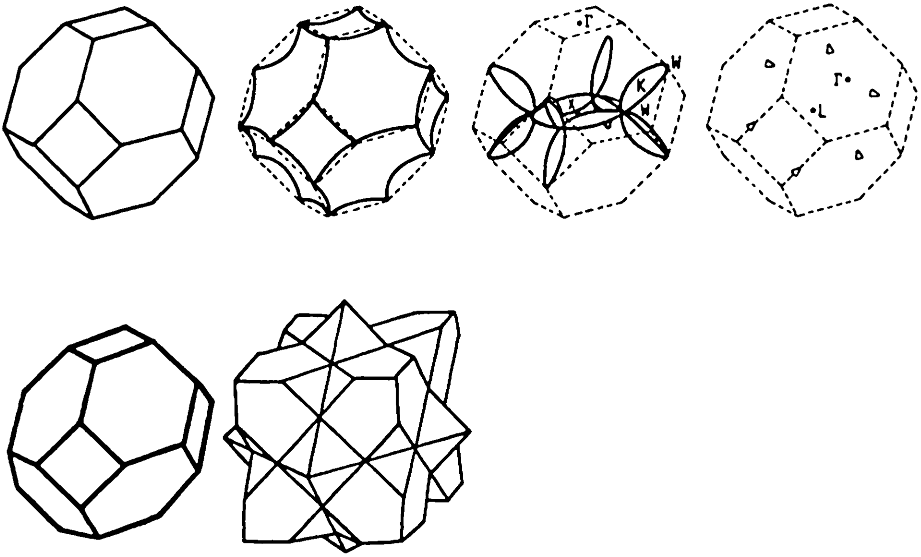


Fig.184.3 The Fermi surfaces of the **empty lattice bands** of aluminum and other valence 3 solids which crystallize in the face-center-cubic lattice. (From W.A.Harrison, **The Fermi Surfaces**, Wiley, 1960.)

190 COMPLETELY FILLED BAND DOES NOT CARRY A CURRENT
(The Concept of Holes.)

In section 173 and Fig.173.1(b) we noted that the valence band of silicon is completely filled by the 5×10^{22} cm^{-3} valence electrons and the conduction band of silicon is empty if no electron-pair bond is broken by external excitation. A completely filled band, such as the Si valence band just described, does not carry electrical current. This can be proved using the E-k diagram we have just developed from the nearly free electron model or the tight-binding model. It is redrawn in Fig.190.1(a) with 8 states in each band and 8 electrons in order to illustrate this conclusion. The velocity of the electrons is the group velocity of the wave packet that represents the moving electron which position and momentum or wave-number have a uncertainty of Δx and Δk respectively and $\Delta x\Delta k\approx 1$ according to Heisenberg's uncertainty principle. The velocity is given by v=group-velocity$=d\omega/dk=\hbar d\omega/\hbar dk=dE/dp=dE/\hbar dk$ where Planck's and de Broglie's relationships are used, $E=\hbar\omega$ and $\lambda=2\pi/k=h/p$ or $p=(h/2\pi)/k=\hbar k$. The dE/dk versus k curve is shown in Fig.190.1(b).

To compute the current, let us label the electrons by subscript j, then the current carried by the electrons in one band is

$$I = -q\sum_{j} v_j = -q\sum_{j} dE_j/\hbar dk_j \tag{190.1}$$

where j=1 to 9 as indicated in Fig.190.1(a) and (b). But electrons 1 and 9 are each counted as 1/2 since they are shared by adjacent k-spaces or Brillouin zones, just like the atoms on the surface of a unit cell are shared by adjacent unit cells. It is obvious from Fig.190.1(b) that currents from electrons 1 and 9, 2 and 8, 3 and 7, or 4 and 6 cancel each other because the velocities of the two electrons in each pair are exactly equal but in opposite directions. For example, electron 2 has a positive velocity while electron 8 has a negative velocity (in the -x direction) and the two velocities are exactly equal in magnitude but opposite in direction. Electron 5 has zero velocity and hence carries no current. Thus, for a filled band, such as the valence band shown in Fig.190.1(a), the total current is zero:

$$I = -q\sum_{j=all} v_j = -q\sum_{j=all} dE_j/\hbar dk_j = 0\ . \tag{190.2}$$

This proves the statement that a filled band carries no current.

Suppose that the j'=6 electron is missing, i.e., the j'=6 state is occupied by a hole as illustrated in Fig.190.1(c). Then, the current carried by the remaining eight valence electrons in the valence band is given by

$$I = -q\sum_{j\neq j'} v_j = \left[-\sum_{j\neq j'} qv_j - qv_{j'} \right] + qv_{j'} \tag{190.3A}$$

$$= \left[-\sum_{\text{all } j} qv_j \right] + qv_{j'} \tag{190.3B}$$

$$= \quad + qv_{j'} \ . \tag{190.4}$$

The sum over all j in the [ ] of (190.3B) is zero since we have just proved in (190.2) that a completely filled band carries no current. (190.4) shows that an unoccupied state in an otherwise completely filled band does carry current. It carries a positive current with a positive charge whose magnitude is equal to the electron charge, q.

The second property of an unoccupied state or a hole is its negative effective mass. This property can be demonstrated using the force equation or the velocity diagram of Figs.190.1(c) and (d). The force equation for an electron is

$$F = dp/dt = d\hbar k/dt \tag{190.5A}$$

or

$$k = k_0 + (F/\hbar)t \tag{190.5B}$$

Thus, k is increasing with time for a positive force ($F_x > 0$) as indicated in Figs.190.1(c) and (d). For the hole at j'=6 in Fig.190.1(c), k decreases since

$$k_h = k_{0h} - (F/\hbar)t. \tag{190.6}$$

Thus, the hole velocity increases with time, indicating that it is a particle with a positive mass. A second proof starts with

$$F = m(dv/dt) = -|m_-|\, dv/dt \tag{190.7}$$

for negative m_- but electron has a negative charge, or

$$F = -qE = -|m_-|(dv/dt) \tag{190.8}$$

Thus,

$$qE = +|m_-|(dv/dt) \tag{190.9}$$

which is just the force equation of a particle with a positive mass $|m_-|$.

Thus, an unoccupied electron state can be treated as a particle with a positive charge and positive effective mass. This is known as a hole.

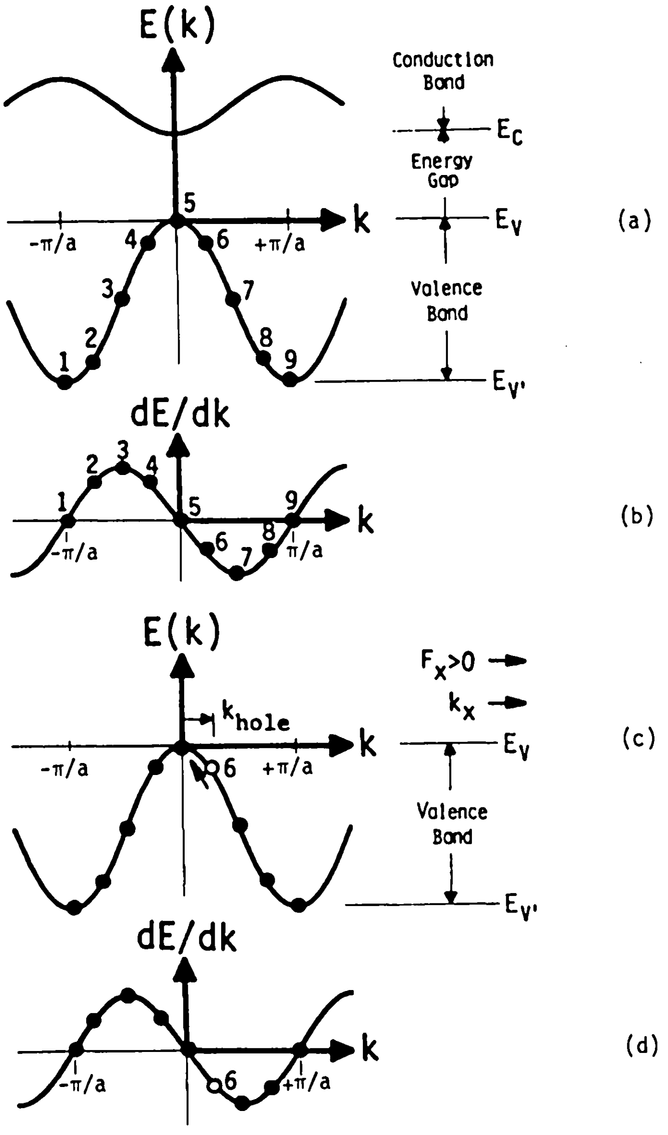

E(k)
Conduction Band
E_C
Energy Gap
E_V
Valence Band
$E_{V'}$
k
$-\pi/a$
$+\pi/a$
1
2
3
4
5
6
7
8
9
(a)
dE/dk
k
$-\pi/a$
π/a
1
2
3
4
5
6
7
8
9
(b)
E(k)
$F_x>0$
k_x
k_{hole}
k
$-\pi/a$
$+\pi/a$
6
E_V
Valence Band
$E_{V'}$
(c)
dE/dk
k
$-\pi/a$
$+\pi/a$
6
(d)

Fig.190.1 (a) The 1-d E-k energy band and (b) the velocity or dE/dk vs k diagrams with 8 electron states and 8 electrons. (c) The 1-d energy (E-k) and (d) velocity (dE/dk-k) diagrams of an unoccupied electron state or a hole and its motion or change of k with time in an electric field $F_x>0$.

199 BIBLIOGRAPHY

The following ten selected historical textbooks and references contain tutorial chapters on the energy band theory of solids which extend the mathematical details of the physics coverage just given in chapter 1. The sequence follows roughly that used by this author when he began to gather materials for teaching undergraduate and graduate level courses on semiconductor physics and devices in 1961. The descriptions in these books are more mathematical than chapter 1 and less historical. Some parts of the selected chapters of these books require a background from a junior or senior course on engineering mathematics in order to follow the detailed derivations. However, the coverage in chapter 1 is sufficient for the students to understand the basic physics and appreciate the importance of the topics described in these books. The references are listed in historical sequence and according to depth starting from the more elementary. They are given in two groups; the first group contains ten selected historical books, and the second group lists the recent textbooks at the senior and first-graduate-year level.

[199.1] William Shockley (Bell Telephone Laboratories, Shockely Transistor Corp., Stanford University), **Electrons and Holes in Semiconductors**, D. Van Nostrand Company, Inc. New York, 1950. Chapter 1: history of transistor electronics revolution and qualitative bond model; chapter 3: experimental evidence of electrons and holes; Chapter 5: Energy band; chapter 6: velocity of electrons; chapter 7: motion in electric field; chapter 14: quantum theory of energy band; chapter 15: quantum theory of velocity.

[199.2] Nevill F. Mott and Harvey Jones, **Properties of Metals and Alloys**, Oxford University Press, London, 1936. Dover paper back edition. This and the following book by Wilson gave the historical first textbook-description of the electron energy band theory of solids and its application to electrical and thermal properties of materials. This book gives less advanced mathematical derivation than Wilson's book.

[199.3] A. H. Wilson (Cambridge University), **The Theory of Metals**, Cambridge University Press, London, 1936, 1958(2nd ed.). Chapter 1 gives historical survey and chapter 2 gives mathematical analyses of the properties one-electron energy bands.

[199.4] Léon Brillouin (College de France, Paris, and IBM-Watson Laboratory), **Wave Propagation in Periodic Structures - Electrical Filters and Crystal Lattices**, Dover Publications, Inc. 1946, second enlarged-corrected edition, 1953. Mathematical solutions are given for many 1-d transmission lines. The Brillouin zone concept is introduced for 3-d crystal lattice structures.

[199.5] Harry Jones (Imperial College, U.London), **The Theory of Brillouin Zones and Electronic States in Crystals**, North-Holland Publishing Company, Amsterdam, 1960. Advanced mathematical foundation including elementary group theory are given for the algebraic and symmetry properties of one electron energy bands in crystals. The Brillouin zone concept is reviewed and the larger Jones zone concept and its geometrical construction are introduced.

[199.6] Frederick Seitz (Princeton, Carnegie Institute of Technology, University of Illinois, National Academy of Science, Rockfeller University), **The Modern Theory of Solids**, McGraw-Hill Book Company, Inc. New York, 1940, and Dover Publications, Inc. New York 1987. This gave the historical first detailed description of the quantum theory of the one-electron energy bands in solid as well as qualitative historical introductions. Mathematical derivation of the one-electron theory for the many-electron solid is given in chapter VI. The approximation methods are described in chapter IX.

[199.7] John C. Slater (M.I.T. and University of Florida), "**The electronic structures of solids,**" in **Handbuch der Physik**, Vol.XIX, pp.1-136, edited by S. Flügge, Springer-Verlag, Berlin, 1956. Professor Slater gave his trade-mark descriptive presentation of the history of the one-electron energy band theories of solids.

[199.8] John R. Reitz (Case Institute of Technology, Case-Western Reserve University), **Methods of One-Electron Theory of Solids**, pp.2-96 in **Solid State Physics**, Vol.1, Edited by Frederick Seitz and David Turnbull, Academic Press Inc. New York, 1955. Gives derivations and many figures on the various historical methods and solutions of the one-electron Schrödinger equation in 3-d crystals. It gives a succinct pedagogical and concise-compact mathematical presentation of those covered in Seitz' and Slater's chapters in [199.6] and [199.7].

[199.9] Joseph Callaway (University of California, Riverside and University of Texas), **Energy Band Theory**, Academic Press, 1964, 1990. Gives a detailed survey of the theory and mathematical methods historically used to calculate the one-electron energy bands in solids with many latest references of journal articles.

[199.10] J.M.Ziman (Cambridge and Bristol Universities), **Principles of the Theory of Solids**, Cambridge University Press, 1964, 1972 second edition. Sums up the essential elements of the one-electron energy band theory and mathematical methods in one short 40-page chapter 3. It also discusses and compares the band and bond pictures in chapter 4 which is one of the first concise description in a textbook.

The following four recent textbooks give the typical textbook presentation of the one-electron energy band theory of solids at the senior and introductory graduate levels. They are more elementary and less mathematical than some of the preceding ten books, and less physical and illustrative than chapter 1.

[199.11] Charles Kittel (University of California, Berkeley), **Introduction to Solid State Physics**, John Wiley & Sons, New York, 1953, 1968-Sixth Edition. Chapter 6 gives the free electron theory of metal using the box model and chapter 7 gives the complex 1-d periodic square-well potential and the 1-d two-plane-wave Fourier expansion analyses of the energy band theory. Tends to be summaritic rather than pedagogical or expository. Earlier editions gave easier treatments.

[199.12] R. A. Smith (University of Sheffield, England, and M.I.T.), **Wave Mechanics of Crystalline Solids**, John Wiley & Sons, New York, 1961. Chapter 1 describes the wave-packet. Chapter 4 gives detailed derivation of the E-k relations of several 1-d examples. Chapters 5 and 8 give general description of electron motion using 3-d energy bands.

[199.13] Neil L. Ashcroft and N. David Mermin (Cornell University), **Solid State Physics**, Holt, Rinehart, and Winston, New York, 1976. Gives probably the best modern treatment of the energy band theory of solids at the intermediate level (senior, first-graduate years). It follows the sequence used by earlier authors (Mott and Jones, and Wilson listed above) but much more comprehensive. Essentially the first fifteen chapters deal with the periodic properties and the one-electron energy band theory.

[199.14] Alexander O.E. Animalu (MIT Lincoln Laboratory), **Intermediate Quantum Theory of Crystalline Solids**, Prentice-Hall, Englewood Cliffs, New Jersey, 1977. Presents the modern theory and methods at about the same level as Ashcroft and Mermin listed above, with more details on some aspects.

199 PROBLEMS

P100.1 Describe the two main physical causes (Answer: Number of particle and interparticle distance or particle density) that make classical Newtonian and classical statistical mechanics inadequate to describe the motion of electrons and atoms in a solid. What hypotheses, laws, principles, models, or methodologies are introduced in quantum mechanics and quantum statistical mechanics to overcome these limitations?

P100.2 What distinguishes hypotheses from laws? When does a hypothesis become a law? Are the following rules hypotheses or laws and why: Newton, Coulomb, Ampere, Planck, de Broglie, Bohr atom, tunneling, Fermi distribution, Bose distribution, Shockley diode, MOSCV, SNS diode, Bethe diode, Mott-Schottky diode, NMOS, CMOS? (Answer those which you already know. Others will be described in detail in the following chapters. Do this problem again at the end of the semester.)

P110.1 What is the single most important fundamental parameter that distinguishes solid, liquid and gas and what are its ramifications? (Hint: A length.)

P111.1 Review the classifications of solid and in what engineering areas are they used?

P120.1 Itemize the reasons why crystallinity and semiconductivity are needed to make transistors.

P131.1 Identify two primitive and two non-primitive cells of the two-dimensional square lattice which are not given in Fig.131.1.

P131.2 Draw the two-dimensional lattice whose primitive translation vector of the lattice is $2a1_x+a1_y$. Show two primitive and two non-primitive unit cells, one of each should be non-rectangular.

P131.3 Identify two primitive and two non-primitive cells of a two-dimensional face-centered square lattice. Give the primitive unit vector and draw it on the figure. Is this a primitive lattice or a composite lattice of two simple lattices?

P132.1 Draw the (100), (110) and (111) planes of a face-centered cubic lattice.

P133.1 Obtain the atomic density expression of the two-dimensional rectangular lattice given in P131.2 using four unit cells similar to those selected for the square lattice in Fig.133.1.

P133.2 Calculate the atomic density of Ga atom, As atom, and GaAs atom-pair in GaAs.

P133.3 What is the concentration of the boron impurity concentration at the middle of a float-zone refined silicon crystal, knowing that k(boron in liquid and solid Si)$=0.8$ and that the Si liquid has 1.0×10^{16} boron/cm^3 initially?

P133.4 If the molten zone is 1 cm during the zone refining growth of a silicon crystal for VLSI application in P133.3, what is the length of the crystal near the seed end that cannot be used if the application requires that the boron concentration must be uniform to within 1%?

P133.5 Oxygen causes indirect problems in controlling the electrical properties of silicon transistors and integrated circuits as described in section 134. The segregation coefficient of oxygen in Si is unity. Can oxygen impurity be removed from Si using the zone refining technique? How can oxygen be removed during the growth of a silicon crystal?

P140.1 Answer the question posed in Problem 110.1 again but this time, give a more detailed and less qualitative or more quantitative but concise (itemized) description of the fundamental reasons.

P141.1 Bohr's first postulate in his hydrogen atom model states that the orbital angular momentum of the bound electron is quantized (or discrete). After (141.4) in the text, a statement is made that this has a very simple geometrical interpretation. Give the graphical demonstration for $n=1$ and $n=4$ (easier to draw than $n=1$) that is anticipated in the text.

P141.2 Are the kinetic and potential energies spatially constant for an electron bound at the n-th orbit with energy E_n in Bohr's hydrogen atom model? Are the answers valid in the quantum-wave model using Schrödinger's equation described in section 156?

P141.3 Draw to scale the electron energy and electron orbit diagram and show the following two optical absorption and emission transitions on the diagrams: the Paschen limit $n=\infty$ <—> $n=3$ for $h\nu=1.51$eV; and $n=3$ <—> unbound for $h\nu=2.100$ eV (yellow light).

P141.4 A ball of 2000 grams is moving at 150meter/sec. [100mile/hour = (100x5280meters)/(3600seconds) = 146.67m/s] A 3000-pound automobile is moving at 65mile/hour. A oxygen molecule in our air is moving at 10^7cm/s. What is the kinetic energy (in joule and in electron-volt), de Broglie wavelength in meters, and Planck frequency in Hz of these moving objects? Which particle may require wave-quantum mechanics to explain its motion and why?

P141.5 A hydrogen molecule is doubly ionized. The inter-proton distance is a. Draw the potential energy of an electron versus x.

P141.6 Show that for a hydrogen atom, $V(r_n) = 2E_n$ and $T(r_n) = -E_n$ where $E_n = -q^4m/[2\hbar^2(4\pi\varepsilon_o)^2n^2]$.

P141.7 Show that $V(r)=-27.2(r_1/r)$eV in a hydrogen atom (r_1=Bohr radius at $n=1$).

P141.8 Draw the potential energy curve of an electron in the field of two motionless protons which are held at $x=-a$ and $x=+a$. Use $r=\infty$ as the reference potential energy. Label also the kinetic energy of the electron if the total energy of the electron is positive, $E>0$.

P141.9 In the two-proton problem above, do you expect the ground state energy of the electron to be more negative (more tightly bound) or more positive than the electron in the one-proton or hydrogen ground state? Why? Label the kinetic energy of the electron bound to the two-proton ground state.

P141.10 The two lowest-energy wavefunctions of the electron bound to the two protons are symmetric and antisymmetric. Sketch and explain why one has a lower energy while the other has a higher energy.

P141.11 An electron is injected into a space between two large-area metal plates which are in parallel and separated by d. A voltage of +10V is applied to the right plate relative to the left plate. What is the potential energy curve of the electron? Assume that an electron at rest on the left metal plate is released. Label the total energy and the kinetic energy of the electron throughout the space between the two metal plates.

P141.12 In the example given in Fig.141.5 there is an applied constant electric field and a proton located at $r=0$. What is the condition assumed for the applied electric field in order to give the specific choice of the reference potential energy used in the figure and text, $V(r=0)=V(x=0)=0$.

P141.13 Can a true bound electron energy state exist in the potential energy given in Fig.141.5? Why? (Hint: No, tunneling.)

P141.14 The electron potential energy is given by $V(x)=V_0(x/a)^2$ where $V_0=10\text{eV}$ and $a=5\text{A}$. Sketch the wavefunctions of the lowest three bound states.

P142.1 Why do we pick the negative sign for p instead of positive sign?

P142.2 Derive the expression for the coefficient A^2 of second order time derivative of (142.0). Show that this violates property (2).

P142.3 Show graphically that the real and imaginary part of the plane wave $\exp[i(kx-\omega t)]$ are traveling waves moving in the positive x direction.

P142.4 Construct a traveling wave moving in the negative x direction.

P142.5 Which of the following are traveling waves and standing waves? Use sketch at $t=0$ and $t=t_1$ to illustrate your answers. (a) $\exp[i(kx-\omega t)]$; (b) $\cos(kx-\omega t)$; (c) $\cos(kx)\cos(\omega t)$; (d) $\exp[-(\alpha x-\omega t)]$; (e) $\exp(-\alpha x)\cos(\omega t)$; (f) $\exp[-\alpha(x-x_0)^2-i\omega t]$. Let $k>0$, $\omega>0$ and $\alpha>0$.

P143.1 Verify the two quantum equations of motion which are defined in Table 143.1 by the quantum mechanical averages: velocity=momentum÷mass and force=rate of momentum change using the definitions of average (143.5) and (143.6). Do this in one-dimension as an introduction and in three-dimension in vector form as a more advanced mathematical-vector-analysis exercise.

P151.1 Refer to Fig.151.1 of the potential energy diagram of electron for electron reflection at a potential step. Show that $T=R=0.5$ when $k_1/k_2=3+2\sqrt{2}=5.828$ or $E/V_0=1.03033$. Why is the solution $k_1/k_2=3-2\sqrt{2}=0.1715729$ not valid?

P152.1 Show that the wavefunction $\psi_2(x)$ also matches $\psi_3(x)$ at $x=a$ in section 152.

P152.2 Show that $\psi_2(x)$ of (152.8) gives the expected result when $k=k_1$ or at zero potential barrier.

P152.3 Show that a repulsive potential causes a phase lag or a time delay while an attractive potential causes a phase lead or a time advance in the problem of resonance scattering by a square potential well of section 152.

P152.4 Sketch the probability amplitude of finding the electron inside the potential barrier of section 152. Find the total probability of finding the electron inside the well. Answer: $a(k+k_1)/2k$.

P152.5 What is a reasonable parameter to use to normalize the probabilities for the problem of resonance scattering by a square well potential given in section 152 and why? Answer: The thickness a.

P152.6 Work out the tedious algebra for a general solution at any electron energy $E > 0$ for the square well potentials given in Figs.152.1(a) and (b). Then, apply the general solution to the repulsive and attractive wells of figures (a) and (b) respectively.

P153.1 For the tunneling through a square potential well problem described in section 153, show that $\partial P_2(y=0)/\partial y=0$ without using the explicit solution of $\psi_2(y)$ given in (153.7).

P153.2 Find the solution $\psi_1(y)=A_1\exp(ikx)+B_1\exp(-ikx)$ and the exact tunneling transmission coefficient of the square potential barrier given in Fig.153.1. Show that the approximate solution is correctly given by (153.6).

P154.1 Derive the tunneling probability of an electron with energy E $(0<E<V_0)$ across a parabolic potential barrier given by three segments: $V(x<0)=0$, $V(0<x<a)=V_0(x/a)_2$ and $V(x>a)=0$. Consider both the incident towards the curved concave wall and the vertical wall. Sketch the position probability functions.

P154.2 Drive the tunneling probability of an electron with energy $E<V_0$ through a convex or bell-shaped parabolic potential barrier given by $V(x)=V_0[1-(x/a)^2]$ with $V_0>0$. Sketch the position probability functions.

P155.1 Derive the equation for the allowed bound state energies of the infinitely deep 1-d well, i.e. with infinitely high wall as illustrated by letting $V_0=\infty$ in Fig.155.1(a). Compute the ground state and first excited state energies (in eV) for a well width of 5A. Compute the well width required to give off visible yellow light at 5000A when the excited electron at n=1 decays down to the ground state, n=0.

P155.2 Complete the algebra that leads to the transcendental equations given by (155.5), (155.6) and (155.7) of the finite 1-d square well.

P155.3 For advanced students, find the condition on (or value of) well depth, V_0, and well width, a, such that only two 1-d bound states exist and that the photon emitted has a wavelength of 5000A (one must be symmetric and the other antisymmetric in order to emit light, why?).

P155.4 Does the square well potential energy persist after one electron is trapped to the ground or an excited state? What is the one-dimensional potential energy expression for the second electron?

P155.5 Describe qualitatively the three lowest possible two-electron bound solutions taking into account electron spin so two electrons can occupy one energy level: both symmetric at the ground state; one at the symmetric ground state and one at the first excited antisymmetric state; and both at the first excited antisymmetric state. Is it possible to have a well width that the first two solutions above do not exist and the lowest allowed energy is one in which both electrons have the one node antisymmetrical wavefunction? Explain.

P155.6 For advanced students, determine mathematically (analytically) if there is a minimum limit on the depth and width of the square well in order to bind the second electron.

P161.1 The experimental electron affinity to an isolated neutral Si atom is 1.39eV while the ionization potential is 8.151eV. In analogy to Figs.161.2 and 161.3, illustrate the electron attachment and detachment transitions at the neutral and negatively charged Si atom by the two **energy level diagrams** and two **transition energy diagrams** or **energy change diagrams**.

P161.2 The electron affinity to an isolated neutral Si atom is 1.39eV while to a neutral Si crystal it is 4.02eV. Why is it larger in Si crystal? Give a simple explanation based on electrostatic or Coulomb force.

P161.3 Draw the electron energy level and transition energy diagrams of a second electron attached to a neutral atom, i.e., a neutral atom can bind two electrons in analogy to Figs.161.1(b) and (c). Use dark lines or curves for the potential energy well that binds the second electron in question. This situation actually occurs at some impurity and defect centers in semiconductors. (Hint: The potential is Coulombic repulsive.)

P161.4 Draw to scale the three electron energy level and transition energy diagrams of the isolated neutral He atom (with two electrons) and the He^+ ion (with one electron) in analogy to Figs.161.1(a)-(c) but taking into account the +2 nuclear charge from the two protons. The first and second ionization potentials of He are 24.587eV and 54.416eV. There is again a solid-state analogy in semiconductors, such as the sulfur (valence=6) impurity center in silicon (valence=4) which will be discussed in chapter 2.

P162.1 The first ionization potential of a neutral hydrogen molecule ($H_2 \rightarrow H_2^+ + e^-$) is experimentally measured as 15.427±0.002eV. Assume that the second ionization potential ($H_2^+ \rightarrow H_2^{++} + e^-$) is 18.08eV and disregard that the two-proton configuration H_2^{++} is unstable. Draw the electron energy level and energy change diagram of this simplest diatomic molecule.

P162.2 Describe the electron configurations denoted by $(1s)^2(2s)^2(2p)^6(3s)^2(3p)^2$, $(1s)^2(2s)^2(2p)^6(3s)^2(3p)$, $(1s)^2(2s)^2(2p)^6(3s)^2(3p)^3$, $(1s)^2(2s)^2(2p)^6(3s)^2(3p)(3d)$, $(1s)^2(2s)^2(2p)^6(3s)^2(3p)(4s)$, and $(1s)^2(2s)^2(2p)^6(3s)^2(3p)^2(3d)$. Give the chemical symbol of each. Use * to denote the excited state.
(Hint: They are all for Si atom in its various ionized, neutral, ground and excited electron states.)

P171.1 Draw to scale the correct 2-dimensional projection of bond diagram of Si in the (111) plane which are hexagons instead of squares shown in Fig.171.2. Use the chemist dot-and-ball model. Stretch out (or develop as the term is used in helix traveling wave tube theory by John R. Pierce) the hexagonal projection so that the projected intercore spacing is the actual intercore spacing, 2.35A. Sketch to scale the sp^3 electron-pair covalent bond orbitals similar to those shown in the inset of Fig.171.2(a) for one hexagon using Table 162.2 to estimate the size of the orbital. Indicate how your estimate is made for the four valence electrons.

P171.2 Repeat problem P171.1 for GaAs. Use filled circle for Ga and unfilled circle for As or vice versa. The lattice parameter or the edge of the cubic unit cell of GaAs is 5.65315A. (For Si, it is 5.43072A so you can compute the interatomic spacing in GaAs using that in Si which is 2.35A.) Use Table 162.2 to estimate the size of the electron orbitals on Ga and As. Indicate how your estimates are made for the three and five valence electrons.

P172.1 Why are the valence electrons the most important on influencing the electrical properties of silicon? Why are the core electrons not as important? Why are the core electrons important to distinguish Si, Ge, GaAs and others?

P172.2 What is the energy of the 1s level in an isolated silicon? (Use the Slater Tables) At what assumed spacing between adjacent atoms would you expect the 1s energy levels to begin to broaden into a band in a highly compressed silicon crystal? How about a 'hydrogen' crystal?

P172.3 What is the valence bond model? What is the energy band model?

P172.4 If all the valence electrons in silicon are in the valence band, what is the minimum photon energy that is just enough to release an electron from silicon into vacuum?

P172.5 What is the photon energy necessary to create an electron-hole pair in GaAs, GaP and Ge?

P172.6 At a finite temperature, many electron-hole pairs are created due to the thermal vibration of the atomic cores. What is the minimum photon energy needed to release an electron from the silicon surface into vacuum?

P172.7 How is the conduction band filled by the valence electrons? Use Cu and its $(4s)^1$ electron as an example. Compare this with Si or C(diamond) and show why Cu is such a good conductor while diamond is not.

P172.8 Draw an energy band diagram, especially near the semiconductor-vacuum interface or the semiconductor surface, in the presence of an electric field.

P172.9 Can we draw current or electrons out of the semiconductor into vacuum by applying an electric field and by what quantum mechanism?

P173.1 How many silicon atoms are there for the energy band diagram shown in Fig.173.2(a). Is the number of levels in the conduction and valence bands drawn consistently and why?

P173.2 How many energy levels are there in the Ge valence band for a Ge crystal of 1.0 cm^3? Repeat this for Si. Answer: $4.41\text{x}10^{22}(\text{Ge/cm}^3)\text{x}1\text{cm}^3 = 4.41\text{x}10^{22}\text{Ge}$. Each primitive diamond unit cell has two Ge atoms (the parallelpiped unit cell of FCC). Thus, there are $2\text{x}2\text{N} = 2\text{x}2\text{x}4.41\text{x}10^{22} = 1.76\text{x}10^{22}$ energy levels.

P173.3 Draw the simplified energy band diagram of GaAs near its surface like Fig.173.2(b). The energy gap is 1.424eV, the electron affinity is 4.07eV at 300K, and the valence bandwidth is about 12eV.

P181.1 As an exercise, derive (181.2) by substituting V(x) from (181.1) into (181.2) and evaluate the integral.

P181.2 Sketch the probability density functions of electrons at energies E_+ and E_- and $k = \pi/a$ and sketch also the periodic potential V(x) using the same origin of the coordinate system.

P181.3 Using the nearly-free electron energy band model obtained in section 181, compute the allowed energy bandwidth for $a = 2.35\text{A}$ (Si interatomic spacing) and show that $2\hbar^2k_1^2/m$ in (181.19) is much larger than the energy gap of Si, $E_G(\text{Si}) = 1.2\text{eV}$, by calculating its value.

P181.4 Derive (181.2) by substituting V(x) from (181.1) into (181.2) and evaluate the integral given by (181.2). This is the Fourier series inversion formula.

P181.5 Using the same Taylor series expansion technique as (181.16), obtain (181.9) and (181.10) following the procedure stated relating to these two equations.

P181.6 Compute the allowed energy bandwidth in the nearly free electron model using $a = 2.35\text{A}$ and show that the term $2\hbar^2k^2/m$ in (181.19) is nearly 4 times the bandwidth and is much larger than the silicon energy gap which is 1.2eV and is the E_G term in (181.19). Compute the effective masses of electrons and holes in units of free electron mass, m, and compare with table values.

P181.7 Show that the effective mass and the energy gap are roughly proportional to each other.

P181.8 Go through the algebra that leads to (181.15) to verify that $\psi_+(x)$ has the higher total energy $E_+ = h^2/8ma^2 + V_1$, and $\psi_-(x)$, the lower total energy $E_- = h^2/8ma^2 - V_1$, if $V_1 > 0$; and the reverse result if $V_1 < 0$. Verify then in a realistic case such as the discussion after (181.15) that the electron concentrated near the atomic or ion core, $\psi_+(x)$, has the lower total energy while electron between the two atomic or ion cores, $\psi_-(x)$, has the higher total energy. The solution for $V_1 > 0$ was postulated by Professor Harry Jones without using the simple physical picture in 1960. He was the first to make the energy-wavefunction correlation using the simple one-dimensional model to illustrate the plane-wave nearly-free electron solution of a periodic potential in the crystal. This was given on p.27, Fig.10 of [199.5]. However, the opposite result was obtained by Professor Kittel on p.127, 2nd-ed. 1956 and p.163, 6th-ed. 1986 [199.11] when considering the larger negative potential energy near the ion core without realizing that it is always valid only for $V_1 < 0$ or the lowest energy gap; and by Professors Ashcroft and Mermin, p.159, 1976 [199.13] in a general mathematical derivation for both $V_1 > 0$ and $V_1 < 0$ without elaborating the physics (i.e. shape of the periodic potential) in a real crystal.

P182.1 Compute the bandwidth estimated by the tight-binding model using the tunneling formula. Let $a = 2.35$A and $W = 10$eV. What is the value of W necessary to obtain the valence bandwidth of Si, 12eV? Sketch the detailed periodic potential for this model Si? Is this value of W reasonable and why? How does the bandwidth vary with interatomic spacing a.

P182.2 Are the dependences of bandwidth on the interatomic spacing, a, consistent among those derived from: (1) the tunneling estimate in the above problem, (2) the tight-binding model (182.7), and (3) the nearly-free electron model in section 181 and problems 181.3 and 181.6?

P182.3 Derive the expression of the electron effective mass near the top and bottom edges of the allowed band using the energy band model derived from the one-dimensional tight-binding approximation (182.7). Are these expressions consistent with those obtained from the nearly free electron model given by (181.19).

P182.4 Find the inverse relationship between the allowed bandwidth and the effective mass using the one-dimensional tight-binding model obtained in (182.7).

P183.1 In analogy to Fig.184.2, expand also the 1-d nearly-free-electron E-k diagram of Fig.181.2(b) to negative values of k. Then, extend to higher energies to show the higher energy gaps. Finally, fold all the higher energy bands back into the reduced k-space from $k = -\pi/a$ to $+\pi/a$ as suggested by the two-band tight-binding model shown in Fig.182.5(a). How would you label the succeeding higher energy nearly-free-electron bands?

P183.2 Verify the result of the translation vector of the reciprocal lattice of the FCC direct lattice given by (184.3).

P183.3 Write out the equation of two or three of the lower empty lattice bands of the FCC lattice shown in Fig.184.2 using (184.1) and (184.3) and verify its configuration degeneracy.

P190.1 Verify mathematically that a complete filled band does not carry current using a one-dimensional atomic chain that has sixteen atoms instead of eight shown in Fig.190.1

P190.2 How fast is k changing with time if the applied electric field is 1MV/cm. $M = 10^6$.

P190.3 If the electron effective mass is $m^* = 0.5m$, what is the velocity of the electron (in cm/s) in an applied electric field of 1MV/cm? Can the solution be obtained using the value of k obtained in the preceding problem and the de Broglie relationship, momentum $= p = \hbar k = (?)\ m^*v$? Why?

Chapter 2
HOMOGENEOUS SEMICONDUCTOR AT EQUILIBRIUM

- **200 INTRODUCTION** 152
  - 201 Homogeneous, 152
  - 202 Equilibrium, 153
- **210 PURE SEMICONDUCTOR CRYSTAL** 159
- **220 IMPURE SEMICONDUCTOR** 160
  - 221 Donors, Acceptors, Isoelectronic Traps, 161
  - 222 Charge States of Donors and Acceptors, 163
  - 223 Binding Energy of Trapped Electrons and Holes, 165
- **230 ELECTRON AND HOLE CONCENTRATIONS AT THERMAL EQUILIBRIUM** 169
  - 231 The Fermi-Dirac Distribution Function, 170
  - 232 Electron and Hole Concentrations (Elementary Analysis), 174
  - 233 Electron and Hole Concentrations (Advanced Analysis), 175
- **240 CALCULATIONS OF THE FERMI ENERGY LEVEL AND THE CONCENTRATION OF ELECTRONS AND HOLES** 181
  - 241 E_F, N and P in Pure Semiconductors, 181
  - 242 E_F, N and P in Impure or Extrinsic Semiconductors, 187
    - * The Mass Action Law, 187
    - * The Charge Neutrality Condition, 188
    - * Criterion of Extrinsic Semiconductor, 190
    - * Components of Carrier Concentration are not Additive, 191
    - * Summary of Carrier Concentration Equations, 192
  - 243 Temperature Dependences of N, P and E_F, 193
  - 244 Intrinsic Temperature, 194
    - * Quantitative Definition of T_i, 195
    - * Components of Carrier Concentration are Additive when $T > T_i$, 197
  - 245 Temperature Dependence of the Electron Distribution, 197
- **250 DEVICE ESSENTIAL ADVANCED TOPICS** 199
  - 251 High Carrier Concentration Effects, 200
  - 252 Impurity Deionization Effects, 203
    - * Conditions of Impurity Deionization, 203
    - * Impurity Occupation Factor, 204
    - * Impurity Deionization Examples, 205
    - ** Deionization at High Impurity Concentration, 205
    - ** Deionization at Low Temperatures, 209
  - 253 Impurity Bands, 214
  - 254 Carrier Screening of Impurity, 218
- **299 BIBLIOGRAPHY AND PROBLEMS** 221

200 INTRODUCTION

The fundamental concepts developed in chapter 1 (electrons, holes, the valence bond, and the energy band) are now used in this and the next chapter, to define the electrical parameters that describe electrons and holes such as the carrier concentration, mobility, and lifetime. These parameters are then used to develop the mathematical model and differential equations in chapter 3. The equations are then used to analyze the electrical characteristics of semiconductors and semiconductor devices in chapter 4 and later chapters. The **homogeneous semiconductor at equilibrium** is treated in this chapter. The properties of electrons and holes in pure crystalline semiconductors are described first, followed by that in the impure crystalline semiconductors. The homogeneity and equilibrium restrictions are removed in chapter 3 in order to analyze the electrical currents in a semiconductor device. We first define the terms, homogeneous and equilibrium, and delineate their underlying physics.

201 Homogeneous

Homogeneous means that the atomic composition of the solid is spatially constant. For example, (i) the distance between adjacent host atoms is a constant in a pure simple cubic crystal, (ii) the unit cell faithfully repeats itself over the entire crystal of a more complex crystal such as diamond or silicon, and (iii) the impurities in a semiconductor are uniformly distributed in the entire crystal. Uniform impurity distribution means that the **local** or **macroscopic concentration** of the impurity does not vary with position in the solid. **Macroscopic** means an average over a sufficiently large sample that contains many **microscopic** fundamental constituents. The sample or the number of constituents to be averaged to give the macroscopic value is chosen to be sufficiently large so that the **root-mean-square deviation** (rms) is sufficiently small to meet the particular experimental or theoretical objective. For example, **concentration**, **local concentration**, or **macroscopic concentration** of an atomic species is defined as (number of atom in a volume element) ÷ (volume of the volume element) = $N/\Delta x\Delta y\Delta z$. In this case, the atom is the microscopic constituent. Thus, to have an acceptably small rms fluctuation in the number of atoms, the volume element, $\Delta x\Delta y\Delta z$, must be large enough in order to contain a large enough number of atoms. If the atomic position is completely random, the fluctuation in number is equal to its square root: $\delta N=\sqrt{N}$ so that the percent fluctuation is $100(\delta N/N)=100/\sqrt{N}$. For a volume containing 10^6 atoms, the number fluctuation is $100/\sqrt{10^6}=0.1\%$. Consider a pure solid whose atomic concentration is typically about $5\text{x}10^{22}$ atom/cm^3. For a cube containing 10^6 host atoms, the volume of the cube is then given by 10^6 atom/($5\text{x}10^{22}$ atom/cm^3) = $(5\text{x}10^{16})^{-1}\text{cm}^3$ = $(271\text{A})^3$ = $(27.1\text{nm})^3$. Thus, the edge of the cube is 271A or 27.1nm which is a very large volume compared with today's micron and submicron Si transistors. The atomic number fluctuation in this volume is then $\delta N=\sqrt{10^6}=10^3$ atoms and the density fluctuation of the host atom is then $[10^3\text{atoms}]/[(5\text{x}10^{16})^{-1}\text{cm}^3]$ = $5\text{x}10^{19}$atom/cm^3 or 0.1% if the pure solid is not crystalline but completely amorphous or with a completely random host atom distribution. As a second example, consider a VLSI grade n-type Si crystal that contains 10^{16}cm^{-3} of randomly distributed phosphorus donor atoms. The cube containing 10^4 phosphorus donors has a volume of $10^4/10^{16} = 10^{-12}\text{cm}^3 = (10^{-4}\text{cm})^3 = 1\mu\text{m}^3$ which will have a phosphorus donor number fluctuation of 1% or a density fluctuation of 10^{14} phosphorus/cm^{-3}. This cube occupies a very large space in the structural body of a submicron Si or GaAs transistor employed in the present and future submicron integrated circuit chips. Thus, random fluctuation in impurity number and density is an increasing serious limitation on manufacturability as transistor dimensions decrease to less than 1 micrometer.

202 Equilibrium

Equilibrium means that the properties under observation do not change with time. Since a material body is composed of electrons and atoms (neutral and charged or ions) which are in continuous motion and which continuously interact or collide with each other due to the electric (Coulomb), magnetic (Ampere), and electromagnetic (photon) forces, there cannot be an equilibrium at the atomic dimension and in the collision time scale. Thus, **microscopic equilibrium is untenable** since the particles' position and velocity undergo continuous changes in space-time and particles are continually generated (detrapped) and annihilated (trapped). However, during an observation or measurement, the average value of a parameter is recorded. This is the **macroscopic average** which is an average not only over space (or over the many particles described above) but also over time (or over the many events such as the collision events). Thus, **macroscopic equilibrium is operationally or phenomenologically definable**. That is, the specific conditions of a macroscopic equilibrium can be defined by the user, based only on certain user-defined requirements, such as accuracy, of a specific user-implemented operation, application, or experiment. It is a **dynamical equilibrium** since the atoms and electrons are in continuous motion. However, the word **dynamic** is frequently omitted in engineering. For example, the term 'thermodynamic equilibrium' used by physicists, chemists and mechanical engineers over a century has been abbreviated as **thermal equilibrium** in transistor theory developed by electrical engineers. This abbreviation is often the cause of an inadequate fundamental understanding of the basics which are needed to design high performance transistors and integrated circuits.

In view of the operational or phenomenological definition, macroscopic equilibria can be delineated into phenomenological groups. Fundamentally based, there are two groups: (i) **electronic equilibrium** (electro-dynamic equilibrium of electrons and holes) more familiar to electrical engineers, and (ii) **atomic equilibrium** (atomi-dynamic equilibrium) more familiar to chemists and materials scientists. Atomic equilibrium is further subdivided into partial, exact and approximate equilibria as indicated in Table 202.1 on the next page.

This delineation into groups is extremely useful not only for a **qualitative understanding** of the fundamental physics of transistor operations but also for the **quantitative analysis** and **design** of transistors and integrated circuits. Such a delineation is crucial since the various phenomena of these groups of partial equilibrium are all very complex and furthermore, they interact in a transistor. Their interactions underlie the design criteria of high performance and high reliability transistors and integrated circuits. In the following paragraphs, the conditions of various **partial** or **individual equilibrium** and the condition of simultaneous equilibria are defined and described with examples. They have a common property: macroscopic equilibrium is the condition of zero net energy transfer, that is, the kinetic energy of the macroscopic unit is constant in time as well as space. This constancy is maintained by the continuous **interparticle collisions**. These collisions cause changes of velocity: effecting exchanges of kinetic energy between particles and between particle species, and smoothing out or homogenizing any macroscopic inhomogeneities (space) and macroscopic variations (time) of the average kinetic energy in space-time.

Table 202.1
Phenomenological Delineation of Equilibrium

ELECTRONIC EQUILIBRIUM
Electron Equilibrium
Hole Equilibrium

ATOMIC EQUILIBRIUM
Atomic Equilibrium (Chemical Equilibrium)
Ionic Equilibrium (Electrochemical Equilibrium)
Thermal Equilibrium (Atomic Vibrations, Heat Transfer)
Mechanical Equilibrium (Many-Atom Massive Body Equilibrium)

Electronic equilibrium is the condition of zero electronic (see below for the ionic component) current and voltage measured from the two contacts or circuit terminals connected to a solid. That is, there is no electrical energy flowing into and out of the solid due to the movements of electrons inside the solid. Not only the net terminal or circuit current is zero but the current from each charge carrier species is also zero. Thus, the electron current and the hole current are each zero. We can have a condition of zero electronic circuit current but a finite measurable (by a voltmeter) voltage across the two terminals of a solid body, such as a semiconductor diode. One example is the open-circuit voltage of a solar cell. In this case, the net terminal current is zero due to open-circuit at the two terminals but the electron and hole currents are individually none zero although they cancel each other exactly. This illuminated and open-circuit solar cell is at an electronic steady-state condition. It is not an equilibrium condition even though the net terminal current is zero. Thus, electronic equilibrium requires simultaneous zero current and voltage.

The first partial equilibrium of the various atomic equilibria is **atom equilibrium**, or equilibrium of uncharged atoms or molecules, generally known as **chemical equilibrium**. The equilibrium condition is similar to that of electronic equilibrium just described except that the electronic current measured by a current meter is replaced by the particle current of the neutral, uncharged atomic particles. Atom equilibrium is the condition that the particle current is zero for each neutral particle species. The particle could be a neutral atom, (H, O, N, ...), a neutral molecule (H_2, O_2, N_2, Cl_2, ...) or a larger molecule. The neutral particle current consists of the diffusion current and the generation-recombination-trapping current. Diffusion is proportional to the concentration gradient, $\partial C/\partial x$, (1-dimensional example). Generation-recombination-trapping is proportional to the rate of change of the particle concentration, $\partial C/\partial t$. Thus, chemical equilibrium means $\partial C/\partial x=0$ and $\partial C/\partial t=0$ or the particle concentration is a space-time constant. For example, a n-type semiconductor is not at an atomic equilibrium condition if its donor impurity species has a spatially varying concentration such as that in a high-temperature-phosphorus-diffused n/p junction diode or n/p/n transistor. However, at room temperatures, the phosphorus particle current due to its concentration gradient is so small that it is zero for all practical purposes. Thus, at room temperature, a n-type semiconductor with a spatially varying phosphorus donor concentration can be considered as at atom or chemical equilibrium. This has an important application in the design of reliable transistors and integrated circuits. Their operating life is rated at 10 years because the phosphorus donor and boron acceptor impurities in a Si n/p junction device do not move significantly at room temperatures.

The condition of **ionic equilibrium** or **electrochemical equilibrium** is identical to that of electronic equilibrium since the translational motion of the ions can be measured by terminal current- and volt-meters. Thus, ion or electrochemical equilibrium is the condition that the electrical current of each ion species is zero. For example, the phosphorus donor ion current, boron acceptor ion current, sodium ion current, and proton current are each individually equal to zero at ionic or electrochemical equilibrium.

Thermal equilibrium denotes the condition of zero heat-transfer. The **heat carrier** is the **propagating atomic vibrational waves** of the oscillating host atoms vibrating about their respective lattice sites, known as **lattice vibration** or **thermal vibration of the lattice**. The atomic vibrational waves are also known as **phonons** when they are quantized using Planck's quantization condition and de Broglie's hypothesis of wavelength-momentum relationship. The vibrations are random due to the perpetual interaction between the atomic nuclei, the core electrons and the valence electrons via the Coulomb force. Thus, thermal equilibrium is the condition of **space-time constancy** of the macroscopically averaged vibrational kinetic energy of the host atoms in the crystal lattice. That is, the thermal current or heat (energy) flux is zero. Heat (energy) is measured by **temperature** which quantifies the average kinetic energy of the random atomic vibrations. It also quantifies the heat transfer rate by conduction via the traveling lattice vibration waves or phonons in a solid. Temperature is defined by $<\text{Kinetic Energy}> = <mv^2/2> = (3/2)k_BT$ where k_B is a universal constant, known as the **Boltzmann constant**. It is equal to the gas constant divided by the Avagadro number and has the numerical value (used by device engineers) of 86.16 μeV/K where $\mu = 10^{-6}$ stands for micro. At 300K (room temperature), $k_BT = 25.85$meV and $(3/2)k_BT = 38.77$meV. These are very small energies compared with the energy gained by an electron falling through a potential drop of 1V which would have an effective electron kinetic temperature of $(1/0.02585)300 = 11{,}605$K. In device analysis and transistor physics, the subscript B in k_B is frequently dropped as long as it is not confused with the wave number, k, of the electron wave.

Thermal equilibrium among two or more species of particles in a solid is attained when the rate of loss and gain of the macroscopically averaged kinetic energy of each species exactly balances so the averaged kinetic energy of all species is the same. This homogenization is effected by energy exchange via the continuous inelastic interparticle collisions. For example, the electrons are in thermal equilibrium with the lattice when $<\text{KE(atoms)}> = <\text{KE(electrons)}>$. Frequently, this is described as '**the electrons are in thermal** (thermodynamic) **equilibrium** with the lattice,' causing some confusion when the words 'with the lattice' are omitted. This omission does not create any ambiguity since the term **thermal** (or thermodynamic) can only mean an equilibrium (or dynamical equilibrium) with the lattice vibration or phonons. However, the fundamental implications may be lost for beginners. As a counter example, the phrase 'the electrons are in equilibrium' means the electrons are in dynamical equilibrium with each other via the frequent electron-electron collisions but they may not be necessarily in dynamical equilibrium with the other particle species present in the solid or in the container. A specific counter example is the Maxwellian distribution of the electrons in Si conduction band in a high electric field. In this case, the electrons are in dynamical equilibrium with each other in order to maintain the Maxwellian distribution in kinetic energy but the electrons are far out of equilibrium with the lattice vibration or phonons.

A convenient description of (dynamical) equilibrium among the particle species is by defining an effective temperature to represent the average kinetic energy of each particle species: $<KE_n> = (3/2)k_BT_n$ for the n-th particle species. Thus, we have T_e (electron), T_h(hole), T_L(Lattice or phonon), T_D(donor ions), T_A(acceptor ions), T_H(proton), etc. At thermal equilibrium, i.e., when all particle species are in (dynamical) equilibrium with each other, we have $T_e = T_h = T_D = T_A = T_H = T_L$.

An applied electric field of a readily attainable intensity (up to about $2\text{x}10^7$V/cm=20MV/cm) is much smaller than the internal electric field from the charged atomic cores (10^8V/cm=100MV/cm) on the periodic lattice of a semiconductor or an electronic solid. (See Fig.141.5.) Thus, the kinetic energy of the host atoms and substitutional impurity ions cannot deviate appreciably from the thermal equilibrium value. If it is a static electric field, the host atoms can only be displaced slightly. If a high-frequency electric field is applied, the additional oscillatory kinetic energy would be quite small at normal field strength. Thus, the lattice temperature of the host atoms and the effective temperatures of the donor and acceptor ions do not deviate significantly from the ambient temperature, that is, $T_L = T_A = T_D$ = ambient temperature. However, electrons, holes and interstitial ions (such as H^+ or proton and Na^+) experience a much smaller internal electric field at their interstitial positions due to screening by the 10^{23} valence electrons. Thus, they can be accelerated to high kinetic energies by an applied electric field of moderate or attainable strength, such as 1-10 MV/cm. Consequently, T_e, T_h, and T_H can easily exceed the lattice or ambient temperature, T_L, in the presence of a moderate electric field, such as 5MV/cm which is easily attainable in a Si p/n junction or a thin SiO_2 film in the VLSI circuit chips. Properties of the semiconductor associated with $T_e >> T_L$ and $T_h >> T_L$ are known as the **hot electron, hot hole or hot carrier effects**. Physically, $T_e >> T_L$ indicates that the electrons are much hotter than the lattice and at a very large deviation from thermal equilibrium with the lattice. However, the electrons may still be in equilibrium with each other by means of frequent electron-electron collisions. In order to describe the electrons by an effective temperature, it is implied that the electrons are in equilibrium with each other through energy exchange via inter-electron collisions.

Table 202.1 also lists **mechanical equilibrium** or **body equilibrium** as one of the partial atomic equilibria. Mechanical equilibrium denotes a motionless solid body in contact with another motionless solid, liquid or gaseous body. It is obviously an operationally defined atomic equilibrium focusing on the motion of the entire solid body. The atomic interdiffusion through the surface contact of two solid bodies, the atomic absorption and evaporation from the exposed surfaces to liquid and gas or vacuum are sufficiently small during the observation period of a mechanical equilibrium that the change of the body shape and weight and the resultant body motion are negligible. However, the thin surface layers are in highly atomic or chemical nonequilibrium states and the properties of an electronic device built on the thin surface layer of semiconductor could undergo significantly changes during the operating life of the device. Thus, mechanical equilibrium is a partial equilibrium in the presence of surface atomic or chemical nonequilibrium.

Having the terms homogeneous, equilibrium, electrical equilibrium, thermal equilibrium, and temperature succinctly defined and extensively illustrated, we now describe the physical properties of the conduction electrons in a solid at a finite temperature under thermal and electrical equilibrium.

In section 173 we described the filling of the electron states or energy levels in an energy band by electrons at absolute temperature. Absolute zero temperature is defined as the frozen lattice in which all the atoms or atomic cores are frozen at their lattice sites, the atomic vibrations ceases, and the lattice vibrational kinetic energy is zero. At $T=0$, the electrons seek the lowest allowed energy or allowed energy levels so that the total energy of the system of particles is at the minimum. Each allowed energy level can only be occupied by two electrons of opposite spin due to Pauli's exclusion principle. Thus, at $T=0$, each of the allowed energy levels will be occupied by two electrons. The allowed energy levels will be filled by two electrons each up to a maximum energy in order to accommodate all the electrons in the crystal. The energy level at this maximum energy is known as the **Fermi level** or **Fermi energy**. All the allowed energy levels above the Fermi energy are unoccupied at $T=0$ since there are no electrons left.

Let us now consider the occupation of the allowed electron energy levels of the solid when the lattice atoms are no longer frozen, represented by a finite temperature, T_L, or nonzero average vibrational kinetic energy. The lattice atoms can be set in vibrational motion if the solid (we consider a solid of a finite volume) is exposed to a beam of electromagnetic radiation or photon in vacuum, or if the solid's surface is exposed to an ambient gas. Energy is transferred to the atoms from these external excitations. In the exposure to light or photons, the electromagnetic field produces the force causing the atoms of the solid to vibrate. In the exposure to a gaseous ambient, the surface atoms are set into vibrational motion by bombardment of the randomly moving gas molecules. The surface atomic vibrations then propagate into the interior of the solid, causing all the interior atoms to vibrate. The net energy transfer via the propagating atomic vibrational waves ceases when the temperature or the average kinetic energy of the vibrational atoms become spatially constant and equal to the ambient temperature. Thermal equilibrium of the solid with the ambient is then reached.

During the approach to atomic thermal equilibrium at T_L described above, the kinetic energy of the electrons is also increased from scattering of the electrons by the vibrating atoms via the Coulomb force. The electrons reach thermal equilibrium with the lattice when the temperature or average kinetic energy of the electrons is equal to those of the lattice or atomic vibrations. This energy transfer moves some of the electrons to the higher energy levels from the lower energy levels which they occupied while at $T=0$. Thus, some of the unoccupied energy levels above the Fermi energy are now occupied and some of the energy levels below the Fermi energy are now unoccupied. The number of electrons as a function of energy at thermal equilibrium is described analytically by a function known as the **equilibrium distribution function** or **Fermi function**, also known as the **Fermi-Dirac distribution function**. It is derived in the following section.

In the above description, the approach to thermal equilibrium is not elucidated in detail. Consider a more detailed (or fine-grain) example on the approach to thermal equilibrium. When a force is applied to a solid, the particle distribution in energy is disturbed or changed. After the force is removed, the distribution will **regress** and eventually **reach** or **approach** the original thermal equilibrium distribution. The approach to thermal equilibrium is effectuated by the continuous collisions among the particles in the solid, and between the particles on the solid surface and the ambient molecules (infinite heat sink). The inelastic collisions among the collisions at the surface transfer the excess kinetic energy, gained by the particles in the solid from the applied disturbance, to the ambient molecules which is the heat sink. If surrounded by vacuum, the excess kinetic energy is radiated by emission of infrared photons. The thermodynamic equilibrium condition is reached when the excess kinetic energy is completely dissipated and the total space-time-averaged kinetic energy of the particles in the solid reaches the original minimum.

210 PURE SEMICONDUCTOR CRYSTAL

A pure semiconductor crystal by definition is one without any impurities and defects. At very low temperatures, very few of the covalent bonds in a pure semiconductor are broken because there is not enough thermal energy or kinetic energy from the thermal vibration of the atomic cores. Thus, there are very few electrons in the conduction band and very few holes in the valence band. Because of the low concentration of electrons and holes, the electrical conductivity of a pure semiconductor at low temperature is very low and the electrical resistivity is very high. For example, the resistivity of a pure silicon at the liquid nitrogen temperature, 77K, is so high ($\sim 10^{40}$ ohm-cm) that it is an insulator with zero conductivity or nearly infinite resistivity for all practical purposes. The best insulator at room temperature (300K) is pure SiO_2 or quartz which has a resistivity of about 10^{18} ohm-cm.

As the temperature is increased by some means that increase the kinetic energy of the ambient molecules and solid atoms, the more rapid vibration of the atomic cores will break some of the electron-pair bonds, releasing a bond electron and leaving a hole behind from each of the broken bond. These electrons and holes are known as the **intrinsic electrons and holes** which are responsible for the **intrinsic conductivity** of the semiconductor. They are termed intrinsic because they are the intrinsic properties of the semiconductor crystal and not due to extrinsic sources such as chemical impurities and physical defects. At room temperatures (about 23C or 300K), many electron-pair bonds are broken by the vigorous lattice vibrations, resulting in about $10^{10}/cm^3$ intrinsic electrons and holes in pure Si and giving a combined conductivity from electrons and holes of 3.2×10^{-6} S/cm or 3.2 μS/cm and a resistivity of 3.2×10^5 ohm-cm or 320 kΩ-cm.

In a pure semiconductor crystal, the electron concentration is equal to the hole concentration at all temperatures because electrons and holes are generated in pairs and they recombine also in pairs. For example, when an electron in the electron-pair or covalent bond is freed from the bond during the generation process, a conduction electron is created between the bonds and a hole is left behind at the bond. The bond diagram, Fig.171.2(a), and the band diagram, Figs.173.2(a) and (b), provide the graphical illustrations.

The general symbol for the electron concentration is N (number/cm^3) where N stands for negative and it is P (number/cm^3) for hole where P stands for positive charge. They are known as the **carrier concentration** or the volume concentration of the **charge carriers**. In a pure semiconductor, they are equal and denoted by the symbol n_i known as the **intrinsic carrier concentration** where the subscript 'i' stands for intrinsic. (n_i of Si is about 10^{10}cm^{-3} at room temperature.) Thus,

$$P = N = n_i. \quad \text{(In pure semiconductor only.)} \tag{210.1}$$

220 IMPURE SEMICONDUCTOR

It is difficult to purify a semiconductor completely to remove all the impurity atoms in the semiconductor crystal. But more importantly, the presence of impurities over a wide range of controlled concentration is the very reason that the conductivity or the resistivity of a semiconductor crystal can be varied by many orders of magnitude, for example, from 10^5 ohm-cm to 10^{-3} ohm-cm in impure Si at 300K. This variation is necessary to build transistors.

The concentration of the impurities can be controlled (i) by impurity doping during crystal growth, (ii) by impurity diffusion into a pure crystal at high temperatures, (iii) by implantation of impurity ions of tens to hundreds keV energy into the surface layer of a crystal with subsequent high temperature heating to remove the ion bombardment damage, (iv) by impurity doping during epitaxial growth of thin semiconductor layers, such as vapor, liquid or solid phase epitaxial and molecular beam epitaxy, and (v) by alloying of an impurity metal or impurity-carrying metal. The precise control of the impurity concentration in a thin layer of 1000 atomic planes of Si is the backbone of the current and future silicon transistor and integrated circuit technology.

Precise control of the impurity concentration in a semiconductor is most important in transistor applications since each impurity atom can give one conduction electron or one conduction hole. Thus, the presence of impurities will control not only the magnitude of the semiconductor conductivity but also the **conductivity type**. Conductivity type means conduction by electron or by hole. **n-type** semiconductor means that electrical conduction or electric current in the semiconductor is mainly carried by electrons. **p-type** semiconductor means that electrical conduction or electric current in the semiconductor is mainly carried by holes. The quantitative definition of the conductivity type is based on the **carrier concentrations** and not the magnitude of the electron and hole conductivities.

Thus, the semiconductor is n-type if N>P, and p-type if P>N. It is intrinsic if $N=P=n_i$.

An impurity atom in a semiconductor can sit at a lattice site where it replaces a host atom or it can sit in the space between the host atoms. When it sits at a lattice site and substitutes for a host atom, it is known as a **substitution impurity**. When it sits in the space between the host atoms (assuming all host atoms are situated at lattice sites), it is known as an **interstitial impurity**. We will focus on the substitution impurities since a simple quantitative picture can be constructed from the bond model to show the emergence of a conduction band electron and valence band hole from a substitution impurity. Interstitial impurity atoms are also present in semiconductor crystal but they are often electrically inactive and hence they do not have as large an effect on the electrical properties of the semiconductor as the substitution impurities. The effects of interstitial impurities are described in advanced device physics and technology books and journal articles.

221 Donors, Acçeptors, and Isoelectronic Traps

There are two main types of substitutional impurities that control the electrical conductivity magnitude and type of a semiconductor, the **donors** and **acceptors**. Elements from column-V of the periodic table (P, As, Sb and Bi) are **donors** when they substitute for a Si host atom. They have five valence electrons per atom while the Si host atom have four valence electrons. Since the four covalent bonds around a Si atom requires only four valence electrons to complete the four bonds in a Si crystal, there is one excess electron from each group V impurity atom. This excess electron will carry electrical current if it is 'donated' to the Si conduction band or released to the space between the Si host atoms and Si covalent bonds. Thus, each group-V donor can contribute one conduction electron. However, not all the group-V impurities are electrically active in a semiconductor to contribute to one electron. For example, electrical activity from the group-V nitrogen impurity has not been observed in Si.

Elements from column-III of the periodic table (B, Al, Ga, In, Tl) are **acceptors** when they substitute for a Si host atom. They have only three valence electrons per atom and are short of one electron per atom to complete the four covalent or electron-pair bonds with the four adjacent Si atoms. This electron deficiency can be made up in three possible ways to complete the four Si covalent bonds: (i) by transferring a valence electron from an adjacent or distant covalent bond, leaving a hole behind at that bond, (ii) by trapping or capturing a conduction electron which is thermally generated via breaking an adjacent or distance covalent bond, leaving a hole behind, or (iii) by trapping or capturing a conduction electron which is released by a donor impurity atom. If there are no donor impurities in the crystal, then only processes (i) or (ii) will operate so that a hole is generated by the presence of an acceptor. The name 'acceptor' comes from the process of accepting an electron to complete the four covalent bonds around the group III acceptor impurity atom. It does not automatically imply that a hole is created or present in the semiconductor, for example, there is no hole in (iii).

The partial periodic table (Table 221.1) summarizes the possible substitutional acceptors and donors in the elemental semiconductor, Si, which are further elaborated next.

Table 221.1
Substitutional Acceptors and Donors
in Elemental Semiconductor, Si

I	II	III	IV	V	VI	VII
Li*	Be*	B	C	N	O	F
Na*	Mg*	Al	Si	P	S	Cl
Cu	**Zn**	**Ga**	Ge	**As**	**Se**	Br
Ag	Cd	In	Sn	Sb	**Te**	I
Au	**Hg**	Tl	Pb	**Bi**	Po	At
Triple Acceptor	Double Acceptor	Single Acceptor	Host Isoelectronic Trap	Single Donor	Double Donor	Triple Donor

Bold: Experimentally confirmed.
* : Interstitial donor impurities.

There is a larger variety of impurities which can act as a donor or an acceptor in compound semiconductors than in elemental semiconductors because there are many different host atoms on the many different lattice sites in a compound semiconductor. Some substitutional impurities can even act as a donor on one lattice site and as an acceptor on the other lattice sites. They are known as **amphoteric impurities**. Other substitutional impurities having the same valence as one of the host atoms are known as **isoelectronic traps**.

Consider the **binary compound semiconductor** GaAs. Ga has a valence of 3 or three valence electrons, thus, impurities with valence greater than 3 will act as a donor when substituting for a Ga. Similarly, As has a valence of 5, thus, impurities with valence less than 5 will act as an acceptor when substituting for As. For example, Si is a donor if it substitutes for a Ga but an acceptor if it substitutes for an As. So Si is an amphoteric impurity in GaAs. Similarly, P(valence=5) and O(valence=6) are donors when they substitute for a Ga(valence=3). When P(valence=5) substitutes for As(valence=5), it is an isoelectronic trap which can trap either an electron or a hole.

When the impurity valence is one unit greater than the host it replaces, it is known as a **single donor**. When the impurity valence is two or three units greater than the host it replaces, such as P and O replacing Ga in GaAs, they are called **double** and **triple donors**. Valence 2 impurities (Zn, Cd, Mg), substituting for Ga, are **single acceptors** while valence-1 impurities (Li, Cu) would be **double acceptors**. Similarly, As in GaAs can be substituted by an impurity to give a single donor (N, 0, S, Se, Te), a single acceptor (Si), as well as double and triple donors and acceptors. In practice not all of the impurities exhibit the expected donor or acceptor electrical activities. The important impurities in GaAs are: Si (amphoteric, i.e. either donor or acceptor), Zn (acceptor) and S (donor), while N(valence=5) is an important and optically interesting isoelectronic trap in GaP when substituting for P(valence=5).

222 Charge States of Donors and Acceptors

Electrical characteristics of a semiconductor device depend on the electrical charge density located at the various parts of the device. Thus, the charge state of the impurities is an important factor in the analysis, design and physics of semiconductor devices.

A group-V donor atom in Si can assume one of the two possible electrical **charge states**. The bond model of the two donors, (a) and (b) in Fig.222.1, shows the two charge states. The fifth or excess valence electron is released from the donor labeled (a) and removed to the space between the bonds. Thus, this donor impurity is not occupied or **unoccupied** by an electron. It is **ionized** with a positive charge of +q when viewed from a distant point at many lattice constants away. This is the **positive charge** state of the donor impurity.

The positively charged donor impurity ion can **capture** or **trap** an electron and become electrically **neutral** when viewed from a distant point. The donor with a bound or trapped electron in a stationary orbit is labeled (b) in Fig.222.1. The neutral donor impurity is said to be **deionized** and is in the **neutral charge state**. The electron is said to be **trapped** by the donor impurity. The **transition** to the **neutral charge state** from the positive charge state, is known as the **impurity deionization process**. Its inverse is known as the **impurity ionization process**.

Similarly, a group-III acceptor can assume one of the two possible charge states in Si, the negative ionized and the neutral acceptor charge states. These are illustrated respectively by the two acceptor atoms labeled (a) and (b) in Fig.222.2. By taking an electron via one of the three mechanisms stated in section 221, the ionized or negatively charged acceptor, (a), has four completed covalent bonds. When a hole is bound or trapped by the negatively charged acceptor, the acceptor becomes neutral. The correct bond picture is shown by the acceptor (b) in Fig.222.2. The common pictures of drawing a half-occupied bond at the acceptor or a hole next to the negatively charged acceptor ion to represent a neutral acceptor are erroneous. In fact, this 'tightly-bound' hole is filled with a valence electron in the cases of B, Al, or Ga acceptors in Si or Ge while the trapped hole has a orbit radius of about 5.5 lattice constants in Si as illustrated graphically, to scale, by the acceptor (b) in Fig.222.2. Calculations of the orbit radius is given in the next section. It is evident that the bond picture of the two charge states of an acceptor impurity given in Fig.222.2 is completely analogous to that of a donor impurity given in Fig.222.1.

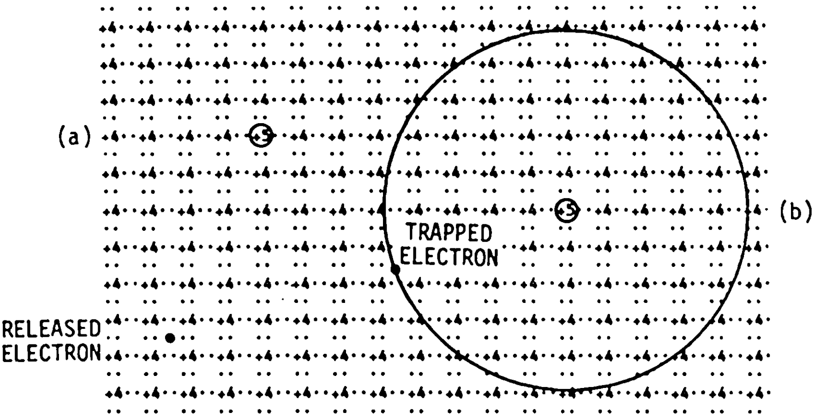


Fig.222.1 The bond picture of the two charge states of a group-V substitutional donor impurity in Si. (a) The ionized and (b) neutral donor configurations.

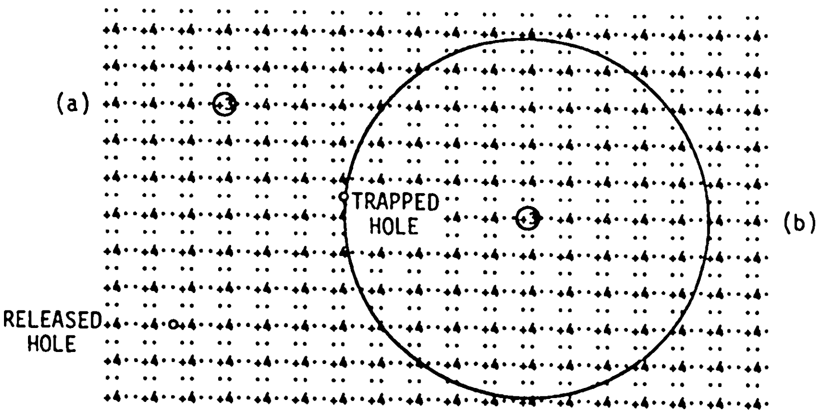


Fig.222.2 The bond picture of the two charge states of a group-III substitutional acceptor impurity in Si. (a) The ionized and (b) neutral acceptor configurations.

223 Binding Energy of Trapped Electrons and Holes

If every one of the phosphorus donors has released its fifth valence electron to the conduction band, the electron concentration would be just equal to the donor impurity concentration. Then $N = N_{DD}$ where N_{DD} (atom/cm^3) is the concentration of the donor impurity atom. N_{DD} includes only those donor atoms situated at the substitution sites. Similarly, the hole concentration would be given by $P = N_{AA}$ if all the acceptor atoms were ionized, that is, all the bound or trapped holes were freed from the respective acceptor atoms.

An excitation must be applied to supply the energy required to release the electron or hole trapped at the attractive Coulomb potential well of the impurity ion core. One form of excitation energy is the kinetic energy of thermal vibration of the host and the impurity atom cores. Another excitation is light or photon. A third is an energetic electron or hole which can knock the trapped electron or trapped hole out of the orbit by impact. These three mechanisms are known respectively as the **thermal**, **optical or photo**, and **impact**. They are described as **excitation**, **emission**, **detrapping**, or **release** of a trapped electron or hole, for example, **thermal emission of a trapped electron** or thermal electron detrapping. They are also known respectively as the **thermal**, **photo**, and **impact ionization** of the impurity. To compute the probability of ionization or detrapping of the three mechanisms, the binding energy of the electron or hole trapped to the impurity must be known.

The simple hydrogen atom model can be used to estimate the binding energy of the electron trapped to a donor and a hole to an acceptor. However, there are two differences between the hydrogen atom in vacuum and the donor or acceptor in a semiconductor crystal. (1) There are many valence electrons and silicon atomic cores between the impurity ion and the captured (trapped or bound) electron or hole. Figure 222.1(b) illustrates this for an electron bound to a positively charged donor ion. Figure 222.2(b) illustrates this for a hole bound to a negatively charged acceptor ion. The large number of valence electrons will screen and hence weaken the attractive Coulomb force exerted by the donor on the bound electron, or the acceptor on the bound hole. The Coulomb force is reduced by the **dielectric constant**. (2) The effective mass of an electron or hole in a semiconductor crystal is different from the free electron mass. This difference arises from the periodic potential which modulates the amplitude of the free electron wavefunction by the amplitude modulation function $u(x)$ in the Bloch function $\psi(x)=u(x)\exp(ikx)$. This was described in the mathematical derivation of the nearly-free electron and tight-binding energy band models given in sections 181 and 182. A third and less important effect is the displacement of the silicon atomic cores surrounding the impurity atom because of the charge difference between the ionized impurity, such as Fig.222.1(a), and neutral impurity, such as Fig.222.1(b). The displacement of the surrounding host atoms is known as the Jahn-Teller effect or lattice distortion which will modify and modulate the force acting on and hence the binding energy of the trapped electron (or hole). It has both a static component as expected from Fig.222.1(a) for the impurity ion and a dynamic component as expected from Fig.222.1(b) for the neutral case which has a trapped electron circulating the impurity ion. Jahn-Teller effect is discussed only in the most advanced course in solid state theory, but it is easily visualized by the bond diagrams shown in Figs.222.1(a) and (b).

The formulae of the electron energy level and radius in the hydrogen atom in vacuum must be modified to take into account the first two crystal effects. It is not possible to simply account for the third crystal or Jahn-Teller effect because many atoms are involved and it is dynamic. The modified Bohr formulae are

$$E_n = -\ m^*q^4/[2\hbar^2(4\pi\varepsilon_s)^2n^2] = 13.605\ (m^*/m)(\varepsilon_0/\varepsilon_s)^2n^{-2}\ \text{(eV)}$$

and

$$a_n = (4\pi\varepsilon_s)\hbar^2n^2/m^*q^2 = 0.5292(m/m^*)(\varepsilon_s/\varepsilon_0)\cdot n^2\ \text{(Angstrom)}.$$

The effective mass of the electron bound to a phosphorus donor impurity in silicon is about 0.45m and the static dielectric constant of silicon is 11.8. Thus, putting $m^*=0.45m$ and $\varepsilon_s/\varepsilon_0=11.8$ in these formulae, the **binding energy** and the radius of the electron trapped at the ground state ($n=1$) of the phosphorus donor impurity are

$$E_1 = -\ 0.044\ \text{eV} = -\ 44\ \text{meV}$$

and

$$a_1 = 13.8\ \text{A}.$$

The effective mass of the hole bound to a boron acceptor impurity in silicon is about $m^*=0.44m$. Using the same static dielectric constant for screening the boron acceptor ion in silicon, $\varepsilon_s=11.8$, then the **binding energy** and the radius of a hole trapped to the ground state ($n=1$) of the boron acceptor impurity are

$$E_1 = -\ 0.045\ \text{eV} = -\ 45\ \text{meV}$$

and

$$a_1 = 14.1\ \text{A}.$$

Two important conclusions can be drawn from these calculations. (1) The binding energy is very small. It is much smaller than the energy gap of silicon, which is 1.18 eV or 1180 meV (milli-electron-volt). It is comparable to the average kinetic energy of the vibrating atomic cores which is $(3/2)kT = (3/2)\text{x}0.026 = 0.039$ eV = 39 meV near room temperature, T = 300K. Thus, we would expect that most of the trapped electrons are freed or released from the donors and similarly, holes from the acceptors at room temperatures. Hence, most of the donor and acceptor impurity atoms are ionized. Because of the small binding energy compared with the energy gap, these group III and V elements are also known as **shallow impurities** since their bound state energy levels are shallow or located near the respective band edges. (2) The orbit radius is large compared with the lattice constant, a, and interatomic spacing, a_0. For Si, a=5.43A and

$a_0=2.53$A. Thus, there are indeed many valence electrons between the impurity ion and the trapped electron or hole. This justifies the use of the static dielectric constant to account for the screening of the Coulomb force which the impurity ion exerts on the bound electron or hole. To verify (2) numerically, we calculate the number of valence electrons enclosed in the sphere whose radius is a_1. There are

$$2.0\times10^{23}(\text{electrons/cm}^3) \times (4\pi a_1^3/3) = 2200 \text{ electrons}$$

where $a_1=13.8$A for an electron trapped to a phosphorus donor in Si, and the Si valence electron concentration, $2.0\times10^{23}\text{cm}^{-3}$, are used.

Two other questions may also be raised which are: (3) the reasons for using the particular values of the effective masses, and (4) the location of these binding energies on the energy band diagrams.

(3) The answer concerning the appropriate effective mass is complicated. Because of the complex three-dimensional energy versus momentum or energy versus wave number, E-k, diagram in a real semiconductor, such as Si illustrated in Fig.183.1 and others in Fig.183.2, a single electron or hole effective mass used in the expression $E=\hbar^2k^2/2m^*$ is inadequate. For each of the many possible electron conduction phenomena (such as drift current, diffusion current, trapping, Hall voltage, photoconductivity, cyclotron resonance, and magneto-resistance) there is a different electron transport effective mass. The effective mass in the binding energy is still different from the transport effective masses. In Si and other indirect semiconductors such as GaP, these electron effective masses are all related to the two fundamental effective masses, m_l and m_t. These are known as the **longitudinal and transverse effective masses.** They characterize the shape of the constant energy surface of the conduction band edge in the E-k diagram, which is not a sphere (given by $E=\hbar^2k^2/2m^*$) but consists of six ellipses, such as $E=\hbar^2(k_x^2+k_y^2)/2m_t + \hbar^2k_z^2/2m_l$, located along the six equivalent <100> directions. The E-k diagram of holes in the valence band of a real semiconductor is even more complicated than that of the conduction band since it is formed from the three 3p atomic wavefunctions of the isolated host atom. As indicated in Fig.183.1 for Si and Fig.183.2 for other semiconductors, the valence band edge is doubly degenerate at k=0 and the constant energy surfaces are warped and represented by the quadratic form equation $E\pm = -Ak^2 \pm \surd[B^2k^4+C^2(k_x^2k_y^2+k_y^2k_z^2+k_z^2k_x^2)]$. The mathematics to obtain the bound state solutions of the Schrödinger equation for these complicated energy-wave number surfaces is very tedious. Thus, in this introductory course, we use a simplified spherical energy surface or E-k diagram in order to illustrate the principle. Consequently, the effective mass of electrons is selected to fit the experimental binding energy of an electron trapped at the phosphorus donor impurity. Similarly, the effective mass of hole is selected to fit the experimental binding energy of a hole trapped at the boron acceptor impurity.

(4) E_I is the binding energy of a electron or hole trapped by the impurity ion, hence it is negative: $E_I = -(E_C - E_D) < 0$ for a trapped electron at a donor impurity, and $E_I = -(E_A - E_V) < 0$ for a trapped hole at an acceptor impurity. The band edge, E_C or E_V, is the reference energy to measure E_I of a trapped electron or hole. They are not measured from vacuum level, $E_{VL} = 0$, which was used as the reference for the periodic potential, because the periodic potential is already contained in the effective mass of these quasi-particles called 'electron' and 'hole' whose kinetic energy is measured from E_C and E_V.

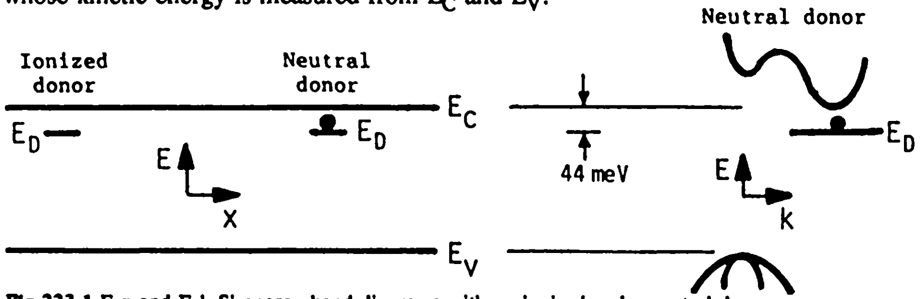

Neutral donor
Ionized donor
Neutral donor
E_C
E_D
E_D
E_D
E
x
44 meV
E
k
E_V

Fig.223.1 E-x and E-k Si energy band diagrams with an ionized and a neutral donor.

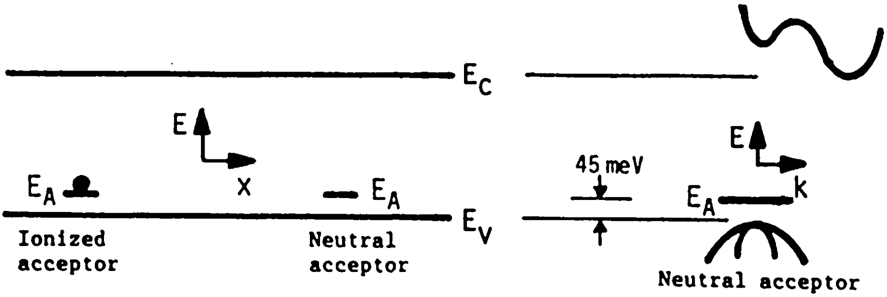

E_C
E
x
E_A
E_A
Ionized acceptor
Neutral acceptor
E_V
45 meV
E
E_A
k
Neutral acceptor

Fig.223.2 E-x and E-k Si energy band diagrams with an ionized and a neutral acceptor.

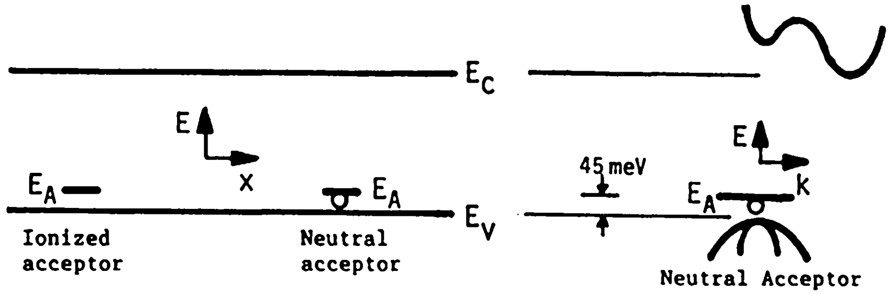

E_C
E
x
E_A
E_A
Ionized acceptor
Neutral acceptor
E_V
45 meV
E
E_A
k
Neutral Acceptor

Fig.223.3 The correct E-x and E-k Si energy band diagrams with an ionized and a neutral acceptor.

The E-x and E-k energy band diagrams of Si containing a donor (phosphorus) and an acceptor (boron) are given in Figs.223.1 and 223.2 respectively, each shows the two charge states of the donor and acceptor. Figure 223.2 is the conventional way of showing the electron occupation of the acceptor energy level, E_A, from the notion that the acceptor has trapped or 'accepted' an electron. This is physically and conceptually incorrect. Figure 223.3 gives the correct diagram that a hole is trapped or freed, which is also consistent with the real real-space bond model of Fig.222.2(b).

The energy level symbols, E_D (D=donor) and E_A (A=acceptor), are the absolute energies measured with respect to the vacuum level or another reference energy level. In semiconductor device and transistor applications, the binding energies, $E_I=-(E_C-E_D)$ of electron and $E_I=-(E_A-E_V)$ of hole, are the relevant energies since the electron and the hole will not conduct if they are bound or trapped. The binding energy is the energy required to free the trapped electron or hole to make it available for conduction of electrical current. Thus, in device physics, we are more interested only in the relative energies or binding energies, E_C-E_D and E_A-E_V, which must be supplied (by heat or lattice vibration) to ionized the donor and acceptor impurities in order to give conducting electron and hole.

230 ELECTRON AND HOLE CONCENTRATIONS AT THERMAL EQUILIBRIUM

We have already shown that for a pure semiconductor, the concentration of electrons and holes are equal and denoted by n_i. It is determined by the intrinsic properties of the semiconductor and n_i is known as the intrinsic carrier concentration. It was also shown that if there are N_{DD} donors/cm^3, then there can be N_{DD} electrons/cm^3 since there is one excess valence electron from each donor atom, provided the excess electron is released from the potential well of the donor ion. Similarly, there would be N_{AA} holes/cm^3 if there are N_{AA} acceptors/cm^3 and every trapped valence hole is released from the potential well of its binding acceptor ion. These released electrons and holes are known the **extrinsic electrons** and **extrinsic holes** since they are not intrinsic to the pure semiconductor itself nor from thermal breakup of the covalent bonds into electron-hole pairs, but instead, depend on some extrinsic substances, such as impurities we have just discussed, or some physical defects.

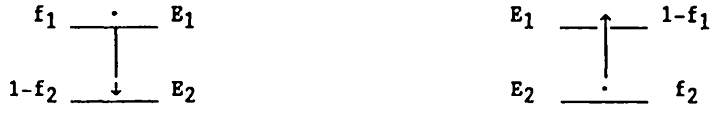
In an impure semiconductor, we need a method to calculate the concentration of the electrons and holes for a given concentration of impurity atom. We cannot get the correct answers if we just do an addition, such as $n=n_i+N_{DD}$ and $p=n_i$ in a semiconductor doped with N_{DD} donors/cm^3, because the hole concentration is decreased below n_i when the electron concentration is increased above n_i. This has been a common mistake made by many textbook authors and practicing engineers. Furthermore, we do not have a method to calculate the value of n_i. Thus, we need to develop some concepts and techniques so that we can calculate the value of n_i at a given temperature and the values of N and P in an impure semiconductor.

We shall first develop the method to compute these concentrations at **thermal equilibrium.** This is the condition where the semiconductor is in contact with a heat bath or heat reservoir which is at temperature T. The ambient surrounding the semiconductor can be taken as the heat reservoir if its temperature is spatially and temporally constant. In order to be strictly at thermal equilibrium, the semiconductor must not be exposed to light, an applied electric or magnetic field, a mechanical force, a chemically active gas or liquid, any atomic and nuclear particle radiations, or a temperature gradient.

231 The Fermi-Dirac Distribution Function

Under this thermal equilibrium condition, the number of electrons in each group of energy levels in the energy range E to E+dE in the conduction and valence bands is a constant, independent of time. What is then the fraction of the electron states or **quantum states** between the energies E and E+dE that are occupied by electrons? We need to know this in order to compute the total number of electrons in the conduction band or holes in the valence band. There are two ways to determine the fraction of energy levels or energy states occupied by electrons: (i) a **fundamental** statistical analysis using the condition that the free energy of the system (the system consists of electrons, host atoms and impurity atoms) is a minimum at thermal equilibrium or the entropy of the system is a maximum, or (ii) a short-cut **non-fundamental** kinetic analysis applied to thermal equilibrium, known as **detailed balance.** We will present the kinetic derivation which is shorter and which also leads to the nonequilibrium concepts encountered in the fabrication and application of a transistor. The kinetic analysis, however, will make use of some results of the fundamental equilibrium statistical analysis based on minimum free energy. Thus, the kinetic analysis is not a first-principle fundamental analysis from scratch, as some teachers and textbook authors were led to believe.

Consider the two energy levels E_1 and E_2 in the conduction band of Si each of which has a large number of electron states clustered about these energies. For example, we can select a sufficiently large volume and a sufficiently wide dE_1 and dE_2 to give many electrons states. We need a large number so that the concept of 'fraction' is statistically meaningful, i.e. with an acceptable fluctuation about the mean. A fraction of the energy states (or energy levels) around E_1, i.e. in the range of E_1 to E_1+dE_1, are occupied by electrons which will be denoted by f_1. Similarly, f_2 denotes a fraction of the energy states around E_2 which are occupied by electrons. At thermal equilibrium, the number of electrons at these two groups of energy states must not change with time. Thus, the number of electrons scattered from states 1 to 2 per unit time by the randomly vibrating atomic cores must be exactly balanced by the number of electrons scattered from state 2 to 1 per unit time. These scattering rates can be computed quantum mechanically if the vibration of the atom is taken into account in the Schrödinger equation by extending the static periodic potential of the crystal, $V(x)$, to a dynamic or time-dependent one, $V(x,t)$. These scattering or **transition processes** are represented by the upward and downward arrows in Figs.231.1(a) and (b). The diagrams show only one electron and two energy levels or states, but in fact, the concept is **macroscopic** and the analysis to follow applies only to a group of many states, energy levels, electrons and transitions, all clustered around ΔE_1 at E_1 and ΔE_2 at E_2, and all located in the volume element $\Delta x \Delta y \Delta z$ and during the time interval Δt. In other words, the analysis is statistical and requires a sample of sufficiently large number of objects (electrons and energy levels) and events (collisions) so that the fluctuation from the mean is sufficiently small as demanded by the required accuracy of a given experiment or application. It is a most common conceptual fault in which the statistical nature of the transition diagrams, such as Figs.231.1(a) and (b) and the results derived from them, are overlooked.



(a) (b)

Fig.231.1 The two electronic transition processes between the two energy states or energy levels, E_1 and E_2, located in a semiconductor volume element dxdydx at a position (x_1,y_1,z_1).

At thermal equilibrium, the number of downward or energy losing scattering transitions (a) must balance the number of the upward or energy gaining scattering transitions (b). This is a macroscopic balance known as the **detailed balance**. It is not a **microscopic balance**: it is not an exact balance of each pair of single electronic upward and downward transition, another cardinal point that is often missed by instructors and textbook authors. Let the macroscopic transition rate coefficients be T_1 at E_1 in Fig.231.1(a) and T_2 at E_2 in Fig.231.1(b), then

$$f_1(1 - f_2)T_1 = \text{downward transition rate in (a)} \quad (231.1)$$
$$= f_2(1 - f_1)T_2 = \text{upward transition rate in (b).} \quad (231.2)$$

The macroscopic transition rate coefficients, T_1 and T_2, are related to each other by the Boltzmann factor: $T_1 = T_2\exp[-(E_1-E_2)/kT]$ where k is the **Boltzmann constant** given by the **gas constant** divided by the **Avagadro number**,

$$\begin{aligned} k &= (\text{Gas constant/Avagadro number}) = R_0/N_A \\ &= 1.3806\times10^{-23}\ \text{J/K} \\ &= 8.616\ \times10^{-5}\ \text{eV/K}. \end{aligned} \quad (231.3)$$

The relationship for T_1 and T_2, involving the Boltzmann factor $\exp(-E/kT)$, is generally taken for granted as the starting point in an elementary or intermediate textbook on semiconductor devices or solid state physics. But, it has a very simple physics origin which can be readily elucidated. It comes from the fact that at a given temperature there are fewer atoms vibrating at higher frequencies, ν, or energies ($E=h\nu$ from Planck's hypothesis) to scatter the electrons. The number of vibrating atoms is inversely proportional to the exponential of its vibrating energy, $\exp(-\beta E)$. This is another assumption, but it is derivable from minimizing the free energy in method (i). Both the exponential dependence and the parameter β can be derived by relating the average kinetic energy of the electrons to the temperature of the solid via thermodynamics or equilibrium statistical mechanics. This gives $\beta=1/kT$ and proves that k is a universal constant. The exponential factor denoted by $\exp(-\beta E) = \exp(-E/kT)$ is known as the **Boltzmann factor** after Boltzmann who first derived it and made use of it in the kinetic theory of gases. Using the Boltzmann factor, the downward and upward transition rates can then be written as $T_1=A\exp(-E_1/kT)$ and $T_2=A\exp(-E_2/kT)$ which gives us $T_1/T_2=\exp[-(E_1-E_2)/kT]$. Note that 'A' is still a macroscopic electronic transition rate obtained from averaging over many microscopic or single events.

Substituting this ratio into the above two equations, we have

$$[(1-f_1)/f_1]\exp(E_1/kT) = [(1-f_2)/f_2]\exp(E_2/kT). \tag{231.4}$$

Since E_1 and E_2 are arbitrarily chosen, that is, the **detailed balance** between the upward and downward transitions must be held for two groups of electrons at any two energy levels, the above equality must be a constant independent of f_1, f_2, E_1 and E_2. It can only depend on the space-time constant temperature, T. Representing this constant by E_F defined by

$$[(1-f_1)/f_1]\exp(E_1/kT) = [(1-f_2)/f_2]\exp(E_2/kT) = \exp(E_F/kT) \tag{231.4A}$$

and solving for f_1 and f_2, we have

$$f_1 = \frac{1}{1+\exp[(E_1-E_F)/kT]}$$

and

$$f_2 = \frac{1}{1+\exp[(E_2-E_F)/kT]}. \tag{231.5}$$

Thus, at temperature T at thermal equilibrium, the fraction of energy states or levels occupied by electron among a group of energy states or levels at an energy E is given by

$$f(E) = \frac{1}{1+\exp[(E-E_F)/kT]} \quad \textbf{(Fermi Distribution Function).} \tag{231.6}$$

This is known as the **Fermi-Dirac (F-D)** or **Fermi distribution function.** E_F is known as the **Fermi Energy** or **Fermi Level.** E_F is the energy where $f=1/2$ or half of the states are occupied by electrons and half are not occupied.

This fraction, f(E), may also be used to denote the probability that an electron state is occupied by one electron or an energy level occupied by two electrons. But it is clear that f(E) is meaningless for single state or single level since it is statistical and is derived from a model of many states or levels clustered around ΔE by selecting a sufficiently large volume element, $\Delta x \Delta y \Delta z$.

The empty levels or states, not occupied by electrons, can be thought of as levels or states occupied by holes. The occupation of the electron states in the valence band by holes is readily illustrated by the bond model. However, it is harder to illustrate the occupation of the electron states in the conduction band by holes using the bond model but it can be readily illustrated using the band model.

The electron energy and temperature dependences of the Fermi distribution function are shown in Figs.231.2(a) and (b). The electron energy is plotted upwards to conform with the convention used in the energy band diagram. Figure (a) shows the fraction of electron states occupied by electron, f, as a function of the electron energy normalized to the thermal energy, kT. The vertical energy axis is shifted so that the energy is measured relative to the Fermi energy, E_F. The Boltzmann approximation, $\exp[-(E-E_F)/kT]$, is shown as broken lines. It is evident that at electron energies greater than several kT, the Fermi function can be accurately approximated by the exponential or Boltzmann function. This is known as the dilute electron gas or nondegenerate approximation. The term 'degeneracy refers to low electron density rather than spin or configuration degeneracy of the electron wavefunction employed in chapter 1.

Figure 231.2(b) shows the temperature dependence of the Fermi function. At absolute zero degree, T=0K, it is a step function abruptly drops to zero when the electron energy is greater than E_F and abruptly rises to unity when $E<E_F$. Thus, all the states below E_F are filled by electrons and all the states above E_F are empty. As the temperature increases, some electrons below E_F are excited thermally to the states above E_F as reflected by the increasingly rounded shoulder and increasingly graduate transition of the Fermi function at energies around E_F. Since E_F is used as the reference energy, the shift of the Fermi function curve with temperature is not explicitly shown in this figure. As we shall see in the following sections, E_F will decrease with lowering temperature in pure silicon or any pure semiconductor when the density of the electron states is higher in the conduction band than in the valence band or the effective mass of electrons is greater than hole. As the temperature is lowered in impure semiconductor, E_F will increase towards the conduction band in a n-type semiconductor and decrease towards the valence band edge in a p-type semiconductor. These variations are necessary in order to maintain electrical neutrality in a homogeneous semiconductor. Detailed

mathematical analyses to give the energy position of the Fermi level will be given in the following sections.

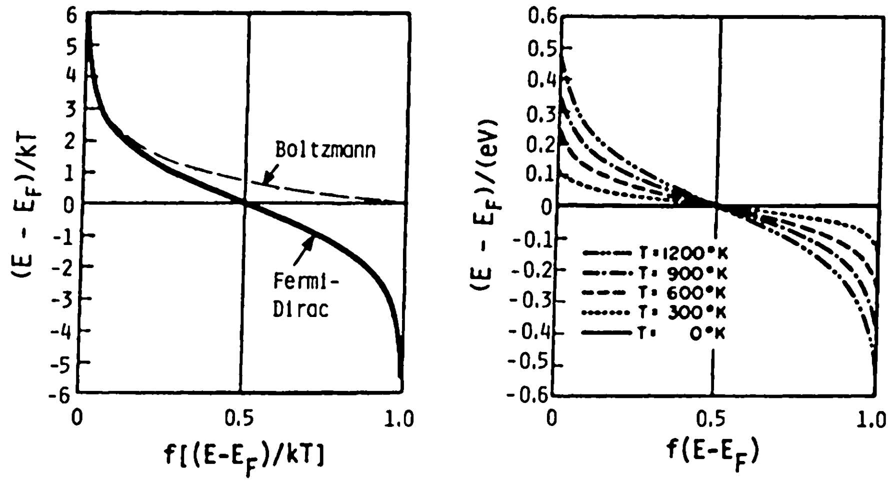


Fig.231.2 The Fermi-Dirac distribution function or the fraction of electron state occupied by electron, f, as a function of the electron energy. (a) The universal solid curve valid at all temperatures and the broken lines in the Boltzmann approximation of low electron densities. (b) The temperature dependence.

232 Electron and Hole Concentrations (Elementary Analysis)

The electron concentration in the bulk region of a semiconductor at thermal equilibrium will be denoted by N_B where the potential energy is taken as the reference. Then, at some other location where the potential energy is $qV(x)$, we would expect the electron concentration to be changed by the Boltzmann factor to take into account the larger or smaller kinetic energy component in the total electron energy, $E = KE + PE = KE - qV(x)$. Thus,

$$N(x) = N_B \exp[+qV(x)/kT] \tag{232.1}$$

where the positive sign shows that the negatively charged electrons prefer locations of more positive electric potentials, $V(x)$. Note that $V(x)$ is the macroscopic electric potential (in units of Volt) seen by a positive charge which gives a negative sign in $-qV(x)$ since the electron charge is negative. In chapter 1, the symbol $V(x)$ and $V(r)$ were used to denote the potential energy (in units of erg, Joule, or electron-volt) of an electron. There should be no confusion since in device analysis only electric and other potential functions are used and the electron potential energy will be denoted by $-qV(x)$ when needed.

Thus, the bulk concentration of the electron is the value at the location $x=x_B$ where $V(x)=V(x_B)=0$. If we adhere to the Coulomb law of the $1/r$ potential, then $V(x=\infty)=0$ and $n(x=\infty)=N_B$. In applications, instead of $x=\infty$, we just say that the electron concentration reaches its bulk value when we are very far away from a specific region of interest where the potential is rapidly changing with position, for example, in the vicinity of a p/n junction or a metal/semiconductor contact.

A similar expression can be written down for the concentration of holes at thermal equilibrium which is

$$P(x) = P_B\exp[-qV(x)/kT] \qquad (232.2)$$

where the negative sign arises from the positive charge of a hole, indicating that the positively charged holes prefer regions of negative electric potential, $V(x) < 0$.

The product of the thermal equilibrium concentration of electrons and holes from (232.1) and (232.2) is then

$$N(x)P(x) = N_BP_B. \qquad (232.3)$$

This is a constant independent of position. One can also show that it is proportional to the deviation from equilibrium if the deviation is small. A basic property of the equilibrium NP product can be verified using the above results in a pure semiconductor where $N(x)=N_B=n_i$ and $P(x)=P_B=n_i$ so

$$N(x)P(x) = N_BP_B= n_i^2. \qquad (232.4)$$

This thermal equilibrium relationship holds even if the semiconductor is not pure and contains impurities which we shall prove in the following section.

233 Electron and Hole Concentrations (Advanced Analysis)

In order to analyze and compute the properties of diodes and transistors, an analytical expression for the intrinsic concentration of electrons and holes n_i must be obtained. This requires the use of two concepts: (i) the occupancy factor or the **Fermi distribution function** of a group of electron states or energy levels at an energy E, and (ii) the volume and energy density of the electronic energy levels or energy states at an energy E, known as the **density of state** or quantum density of state. The product of these two factors gives the electron (or hole) **volume density** or **concentration** in an energy range between E and E+dE.

The total electron concentration in the conduction band is then obtained by summing up all the electrons occupying the electron energy states in the conduction band. This summation can be replaced by an integral since the crystal is large and contains many electrons and many energy levels as indicated by Fig.173.1(b). The concentration of the electrons in the conduction band is then

$$N = \int_{E_C}^{E_{C'}} f(E)D(E)dE. \qquad (233.1)$$

D(E)dE is the number of electron states per unit volume of crystal between the energy E and E+dE in the conduction band. This volume density of electron states can be obtained by counting the number of states in the conduction band (or the valence band) using the energy-wave number diagram. For example, let us use the 1-d in Fig.190.1(a). We have 8 states from $k=-\pi/a$ to $+\pi/a$ for $N=4$ atoms in the 1-d cyclic cell. Thus, for a 1-d crystal of N atoms or a length of $L=Na$, there are 2N states in one band. Since the length of k for a band is $2\pi/a$, as indicated in Fig.190.1(a), the number of states in a range dk between k and k+dk is

$$[(dk)/(2\pi/a)]2N = 2(dk/2\pi)(Na). \qquad (233.2)$$

The number of states in dk per unit length of crystal, $L=Na$, is then

$$D_1(k)dk = 2(dk/2\pi)(Na)/L = 2(dk/2\pi). \qquad (233.3)$$

The factor 2 comes from the two electron spins, that is, for each energy level designated by an allowed k, there are two electron spin states, one spin up and the other spin down. The number of states in the energy range dE per unit length of a 1-d crystal contains an additional factor of 2 because in each dE there is a dk at +k and a dk at -k. Thus,

$$D_1(E)dE = 2D_1(k)dk = 2\cdot 2(dk/2\pi). \qquad (233.3A)$$

For 3-d, the volume density of states in the range of $dk_x dk_y dk_z$ is then,

$$D_3(k)d^3k = 2(dk_x/2\pi)(dk_y/2\pi)(dk_z/2\pi) = 2[4\pi k^2 dk/(2\pi)^3]. \qquad (233.4)$$

To get $D_3(E)dE$, (233.4) is integrated over a thin shell between the two constant energy surfaces E and E+dE in the 3-d E-k space. Assuming a spherical energy surface, $E-E_C=\hbar^2(k_x^2+k_y^2+k_z^2)/2m^*$, with constant m^*, then D(E)dE in (233.1) is

$$D_3(E)dE = \frac{d}{dE}\iiint D_3(k)dk_x dk_y dk_z \cdot dE \text{ (Integrated between E and E+dE.)}$$

$$= [(2m^*)^{3/2}/2\pi^2\hbar^3]\sqrt{E - E_C}\,dE \qquad (233.5)$$

where $\hbar=h/2\pi$. The function $D_3(E)$ or D(E) is known as the volume-energy density of electron states (in the unit of number/cm^3-eV) or density of states. It appears in many semiconductor physics problems. $D_3(E)$ in (233.5) for the 3-d crystal can be substituted into the integral for N, (233.1). The integral can then be evaluated to give an approximate analytical expression for the electron concentration, N. This is worked out in the following paragraph.

For low electron concentration in the conduction band, commonly known as the **nondegenerate statistics** or **nondegenerate distribution**, the Fermi-Dirac distribution can be simplified because the exponential factor in the denominator is much greater than unity. Thus,

$$f(E) = \{1 + \exp[(E_F-E)/kT]\}^{-1} \approx \exp[-(E_F-E)/kT]. \qquad (233.6)$$

This was graphically illustrated by the broken lines in Fig.231.2(a). The electron concentration integral (233.1) then becomes

$$N = \int_{E_C}^{E_{C'}} f(E)D(E)dE$$

$$= [(2m^*)^{3/2}/2\pi^2\hbar^3]\cdot\left[\int_{E_C}^{E_{C'}} \frac{\sqrt{E - E_C}dE}{\{1 + \exp[(E-E_F)/kT]\}}\right] \qquad (233.7)$$

$$\approx [(2m^*kT)^{3/2}/2\pi^2\hbar^3]\cdot\left[\int_0^{\infty} \sqrt{\varepsilon}d\varepsilon\ \exp(-\varepsilon/kT)\right]\cdot\exp[-(E_C-E_F)/kT] \qquad (233.7A)$$

This approximation, (233.7A), can be evaluated to give

$$N = N_C\cdot\exp[-(E_C-E_F)/kT]. \qquad (233.8)$$

where

$$N_C = 2(m_ekT/2\pi\hbar^2)^{3/2} = 2(2\pi m_ekT/h^2)^{3/2} \qquad (233.8A)$$

$$= 2.50\times10^{19}(m_e/m)^{3/2}(T/300)^{3/2}\ cm^{-3}. \qquad (233.8B)$$

N_C is known as the **effective density of state in the conduction band**, first introduced by Shockley. It gives the correct electron concentration if all the conduction band states are clusters at E_C at a concentration of N_C. In (233.8A) and (233.8B), the density-of-state effective mass, m^*, is denoted by m_e for the electrons in the conduction band.

The arithematics leading to (233.8) from (233.7) are as follows. In (233.7A), the integration variable E was changed to $\varepsilon = E - E_C$. The upper limit of the integration, $E_{C'}$, was approximated by infinity because the conduction bandwidth is much larger than kT as indicated by CB-width$=E_{C'}-E_C=4eV >> kT=0.026eV$), thus, $(E_{C'}-E)/kT=160 \approx \infty >> 1$. The two approximations, Boltzmann and $E_{C'}=\infty$, are graphically illustrated in Figs.233.1(a)-(c). D(E) is shown in Fig.233.1(a), f(E) in (b), and f(E)D(E) in (c). It is evident that when $E_F << E_C$ in the energy gap, f(E) has decayed to the exponential function in the conduction band where D(E) is not zero. And, the integrand f(E)D(E) given in Fig.233.1(c) shows that most of the electrons or the major contribution to the electron density integral comes from an energy range of about 5kT above E_C. To evaluate the integral of (233.7A) analytically, the following integral is used which is looked up from an integral table, such as Pierce.

$$\int_0^\infty \exp(-y)\sqrt{y}dy = \sqrt{\pi}/2 \qquad (233.8C)$$

Fig.233.1 (a) The density of state, D(E), (b) the Fermi function f(E), and (c) the electron density per unit energy range, f(E)D(E), in the conduction band. ↓↓↓ (see below)

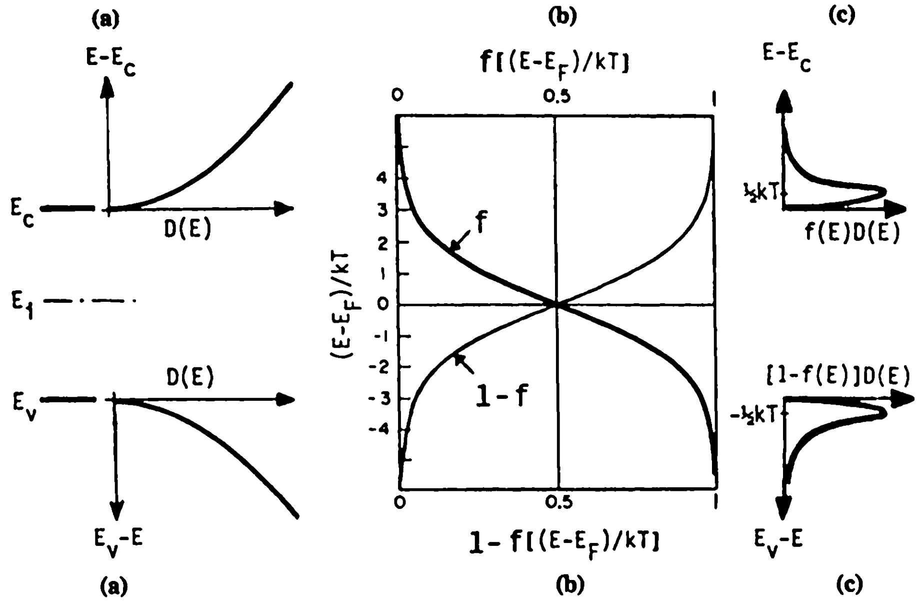


Fig.233.2 (a) The density of state, D(E), (b) the complement of the Fermi function or the hole distribution function, 1-f(E), and (c) the hole density per unit energy range, [1-f(E)]D(E), in the valence band. ↑↑↑ (see above)

Equation (233.8) shows that the electron concentration is exponentially dependent on the Fermi energy measured from the conduction band edge, E_F-E_C. This is known as the Boltzmann approximation. It is valid only when the Fermi function can be accurately approximated by the Boltzmann or exponential function. The approximation is good when the Fermi level is inside the energy gap and at least 2kT away from the conduction and valence band edges which corresponds to a low electron concentration (less than about 10^{18} cm^{-3}). For high electron concentration, such as that encountered in conductors and metals, the full Fermi function must be used and the integral is known as the Fermi-Dirac integral of the one-half order. This is discussed in sections 25n on heavily doped semiconductors and metals. (See also problems in section 299.)

A similar derivation gives the hole concentration in the valence band. It is just the number of states in the valence band not occupied by electrons as discussed in section 190. Figures 233.2(a)-(c) show the components of the integrand of the hole concentration which are just the mirror image of those of electrons given in Figs.223.1(a)-(c). In the Boltzmann approximation, the hole concentration is

$$P = \int_{E_{V'}}^{E_V} [1-f(E)]D(E)dE \tag{233.9}$$

$$= N_V \exp[-(E_F-E_V)/kT] \tag{233.10}$$

where

$$N_V = 2(m_h kT/2\pi\hbar^2)^{3/2} = 2(2\pi m_h kT/h^2)^{3/2} \tag{233.10A}$$

$$= 2.50\text{x}10^{19}(m_h/m)^{3/2}(T/300)^{3/2}\ cm^{-3}. \tag{233.10B}$$

From measured fundamental effective masses of the constant energy surfaces of the conduction and valence bands using cyclotron resonance absorption experiments at (4.2K) and lower temperatures, $m_e=1.065m$ and $m_h=0.647m$ were calculated for Si. Thus, at room temperature or T=300K, the effective density of states in the Si conduction band is $N_C=2.75\text{x}10^{19}$ cm^{-3}. In the Si valence band, it is $N_V=1.30\text{x}10^{19}$ cm^{-3}. Similar measurements for GaAs gave $m_e=0.067m$, and $m_h=0.47m$, and at T=300K, $N_C=4.33\text{x}10^{17}cm^{-3}$ and $N_V=8.05\text{x}10^{18}cm^{-3}$.

The dependences of the integrand of the electron and hole concentrations, N and P, on the position of the Fermi energy, E_F, are shown in Figs.233.3(a) and (b). The integrands at two Fermi energies, E_{F1} and E_{F2}, are plotted. It is evident that the electron concentration increases and the hole concentration decreases as E_F moves up towards the conduction band edge from the mid energy gap or midgap position. The reverse is true when E_F moves down towards the valence band edge from the midgap position.

Fundamental considerations of the Fermi distribution function and analytical justifications of the Boltzmann approximation are further elaborated in Problems 233.n given at the end of this chapter. In sections 25n, the analysis is also extended to high electron and impurity concentrations for metals and degenerate semiconductors in which the Boltzmann approximation is no longer valid. These are practically important materials, for example, the emitter and base layers of submicron picosecond bipolar junction transistors have very high concentrations of electrons and holes.

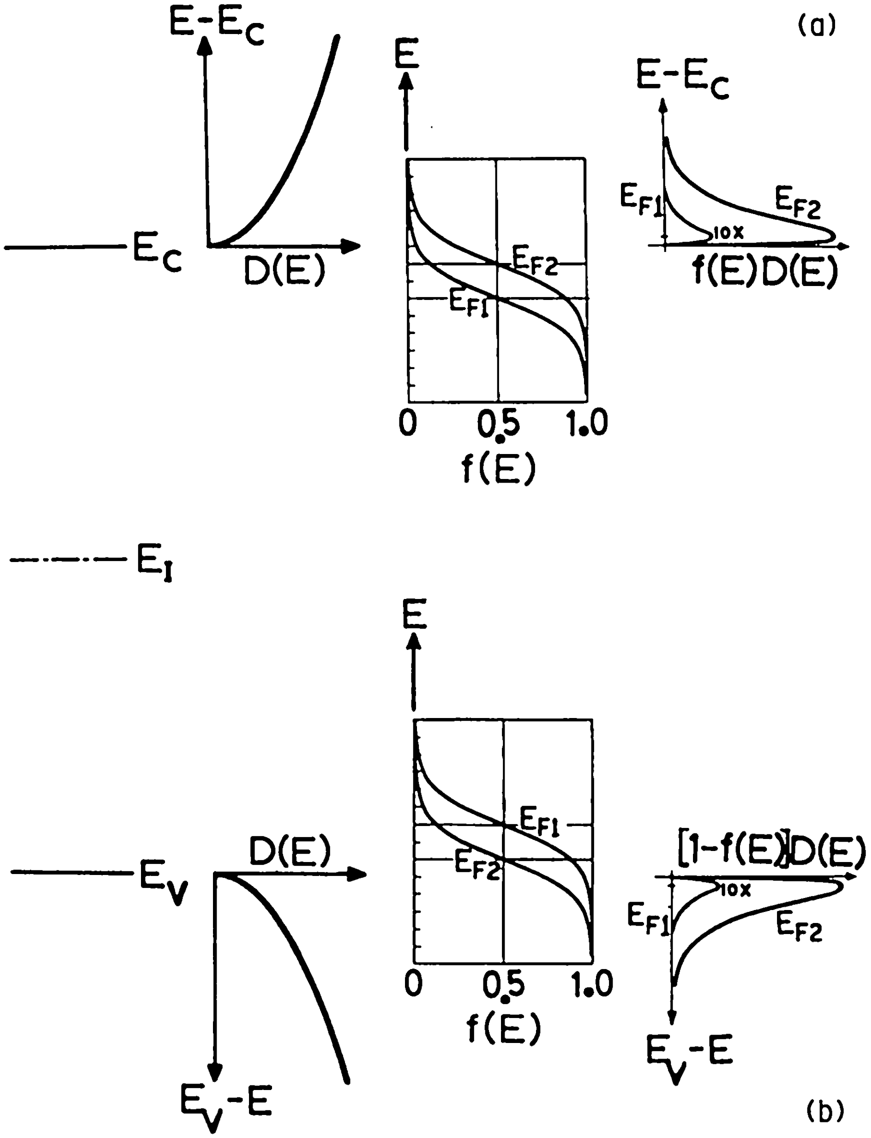

(a)
E-E_C
E_C
D(E)
E
E_F2
E_F1
0
0.5
1.0
f(E)
E-E_C
E_F1
E_F2
10X
f(E)D(E)
E_I
E
E_F1
E_F2
0
0.5
1.0
f(E)
E_V
D(E)
E_V-E
[1-f(E)]D(E)
10X
E_F1
E_F2
E_V-E
(b)

Fig.233.3 Variation of the carrier concentration integrand and concentration with the position of the Fermi energy, E_F. (a) electron, N, and (b) hole, P.

240 CALCULATIONS OF THE FERMI ENERGY LEVEL AND THE CONCENTRATION OF ELECTRONS AND HOLES

The procedures to calculate the Fermi level position and the electron and hole concentrations in pure and impure semiconductors are now described. An intrinsic temperature is then defined.

241 E_F, N and P in Pure Semiconductors

From the electron and hole concentrations, (233.8) and (233.10), the position of the intrinsic Fermi level and the magnitude of the intrinsic carrier concentration, n_i, are determined as follows. The NP product gives

$$n_i = \sqrt{NP} = \sqrt{N_C N_V \exp[-(E_C-E_V)/kT]} = \sqrt{N_C N_V}\exp(-E_G/2kT). \qquad (241.1)$$

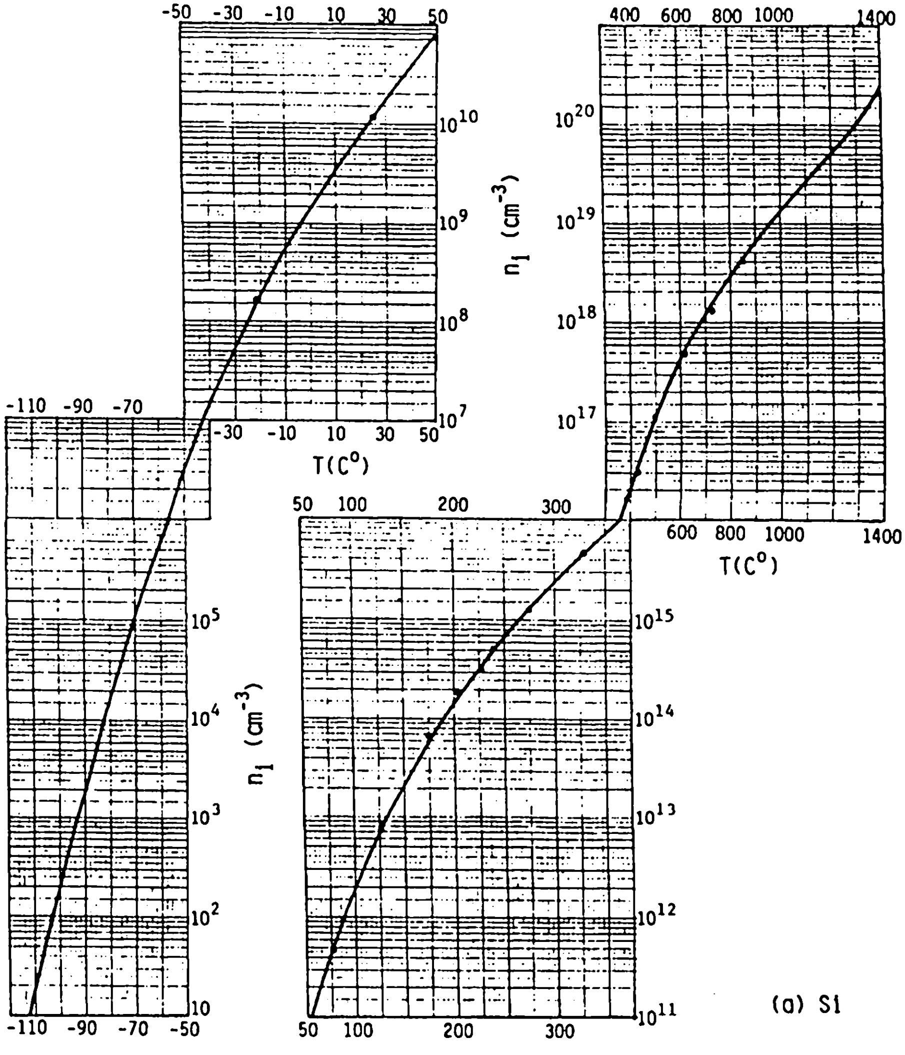
Plug in N_C, N_V and E_G, then $n_i = 10^{10}$ carrier/cm^3 for Si and 10^6 carrier/cm^3 for GaAs at $T=300K$. Very accurate values of E_G must be used to give an acceptable accuracy in n_i because $E_G/2kT \approx 23$ is large and $\exp(-E_G/2kT)$ or n_i would be in error by 10% when E_G is in error by only $0.2kT = 0.2 \times 25.85$meV $=$ 5meV.

Using $N=P$, (241.1), and (233.8) or (233.10), the **intrinsic Fermi level**, denoted by E_I, is given by

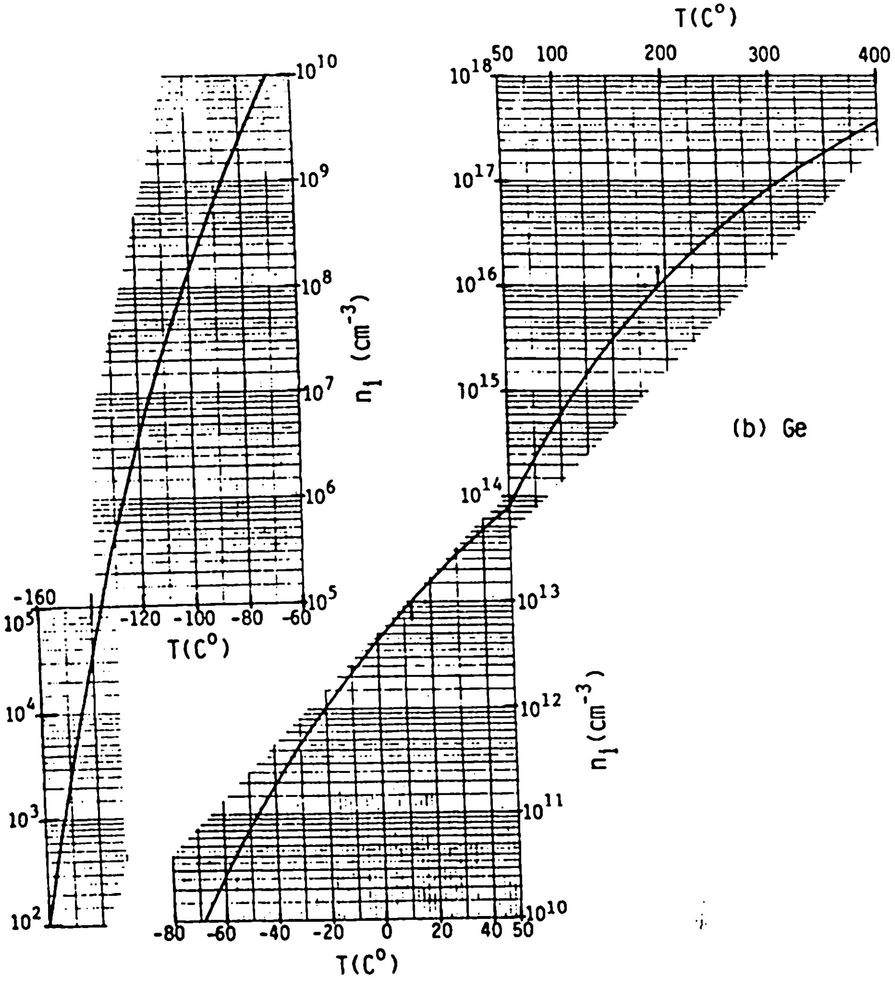
$$E_I = E_F(\text{when } N=P=n_i) = (1/2)[(E_C + E_V) + kT\log_e(N_V/N_C)]. \qquad (241.2)$$

This shows that E_I lies slightly below the middle of the energy gap, or the midgap, $E_{MG} = (E_C + E_V)/2$, because $\log_e(N_V/N_C)<0$ in (241.2) owing to $N_C > N_V$ in Si due to the larger density-of-state effective mass of electron than hole, $m_e > m_h$. In many applications and discussions of results, the midgap energy, E_{MG}, is often not distinguished from the intrinsic Fermi level, E_I, since they nearly coincide as indicated by (241.2). But for numerical accuracy, one must use the intrinsic Fermi level instead of the midgap energy to compute N and P. [See (242.3) and (242.4)].

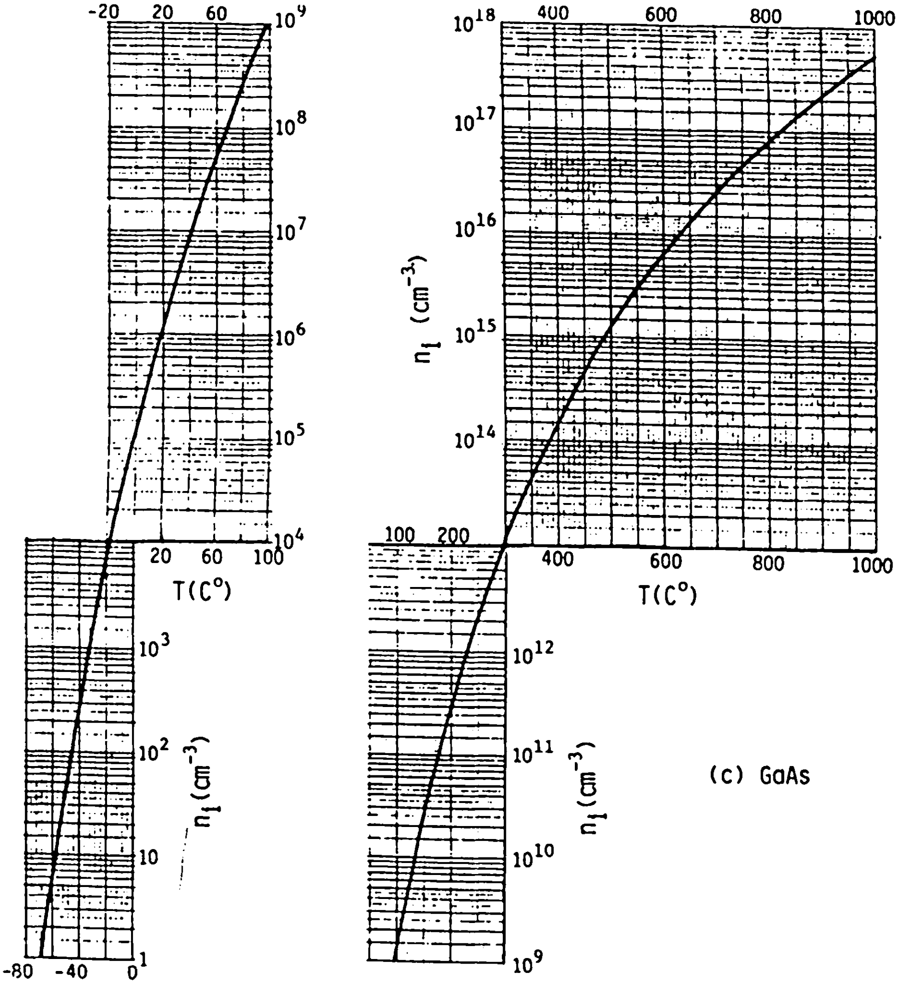
The experimental intrinsic carrier concentration in Si, Ge, GaAs and GaP are plotted as a function of temperature in Figs.240.1(a)-(d). They vary with temperature over many orders of magnitude. The variation is larger and n_i is smaller in materials with larger energy gap. This trend is summarized in Fig.240.1(e) which gives $\log_{10}n_i$ versus the reciprocal temperature, 1000/T(K), showing the consistency with the energy gap trend Ge(0.74eV) < Si(1.17eV) < GaAs(1.52eV) < GaP(2.34eV). Such a semilog plot is known as the **Arrhenius plot.** The lines are nearly straight lines. The slope is known as the thermal activation energy, which is $E_G/2$ if n_i is divided by $T^{3/2}$.



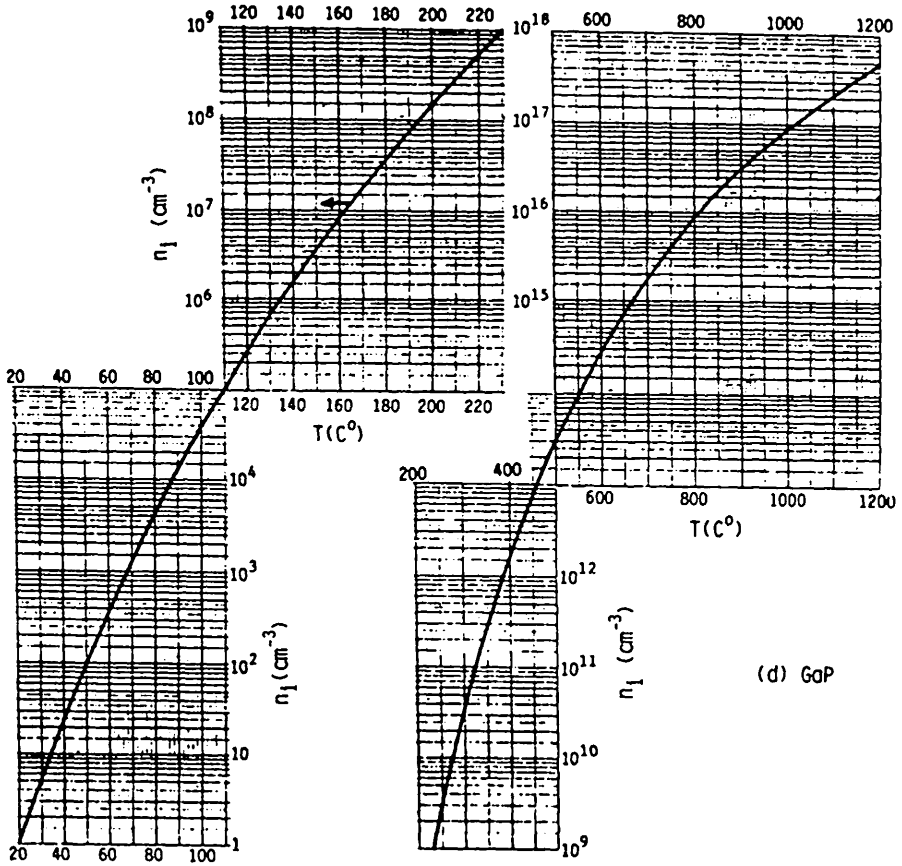
Figs.241.1 The intrinsic carrier concentration as a function of sample temperature in (a) Si, (b) Ge, (c) GaAs, and (d) GaP. (e) Arrhenius plot of n_i versus reciprocal temperature, 1000/T(K), for Si, Ge, GaAs and GaP.



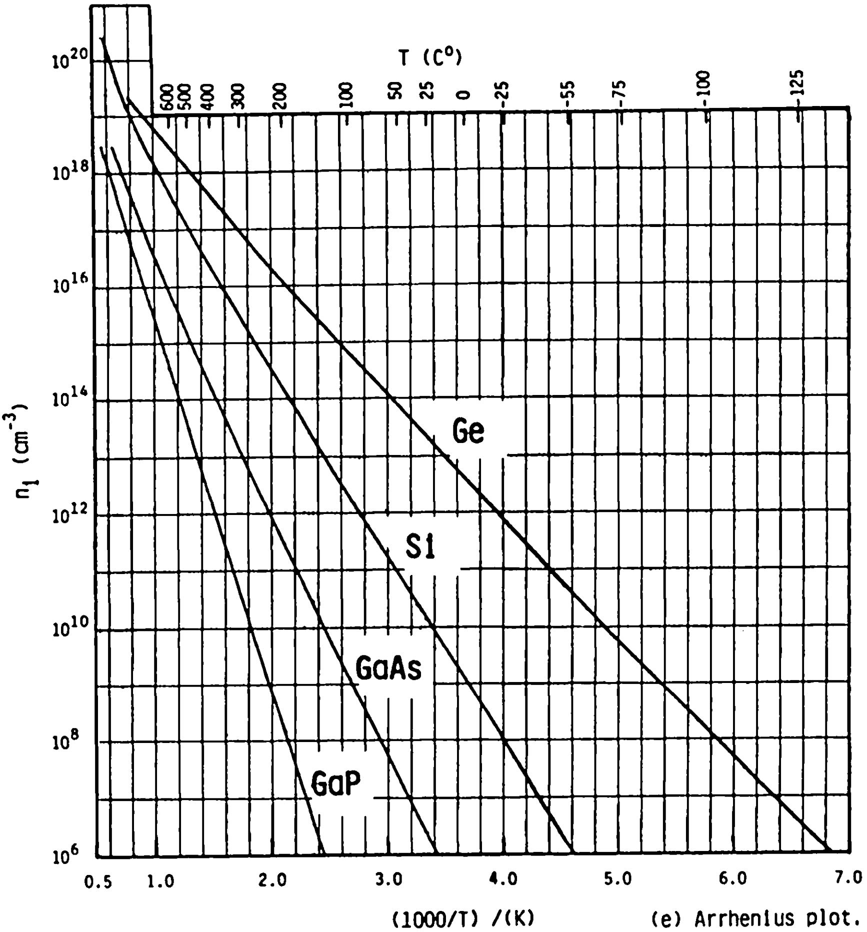
Figs.241.1 The intrinsic carrier concentration as a function of sample temperature in (a) Si, (b) Ge, (c) GaAs, and (d) GaP. (e) Arrhenius plot of n_i versus reciprocal temperature, 1000/T(K), for Si, Ge, GaAs and GaP.



Figs.241.1 The intrinsic carrier concentration as a function of sample temperature in (a) Si, (b) Ge, (c) GaAs, and (d) GaP. (e) Arrhenius plot of n_i versus reciprocal temperature, 1000/T(K), for Si, Ge, GaAs and GaP.



Figs.241.1 The intrinsic carrier concentration as a function of sample temperature in (a) Si, (b) Ge, (c) GaAs, and (d) GaP. (e) Arrhenius plot of n_i versus reciprocal temperature, 1000/T(K), for Si, Ge, GaAs and GaP.



Figs.241.1 The intrinsic carrier concentration as a function of sample temperature in (a) Si, (b) Ge, (c) GaAs, and (d) GaP. (e) Arrhenius plot of n_i versus reciprocal temperature, 1000/T(K), for Si, Ge, GaAs and GaP.

242 E_F, N and P in Impure or Extrinsic Semiconductors

In an impure semiconductor whose impurity concentration is either spatially constant or changing very slowly with position, two conditions must be used to obtain the two unknowns, the electron and hole concentrations. The given parameters are: (i) the temperature, (ii) the intrinsic carrier concentration, n_i, which can also be computed at the given temperature if it is not given and if the energy gap and the effective masses are accurately known, and (iii) the concentration of the donor (N_{DD}) and the acceptor (N_{AA}) impurities. The two conditions are: (1) the mass action law and (2) the charge neutrality condition.

(1) The Mass Action Law

The mass action law is the electronic analogue of the chemical reaction in which the reacting species are the electrons and holes and the reactions are the electron-hole recombination and generation processes. The electronic 'chemical' equilibrium equation is then given by

electron + hole <————> perfect covalent bond.

The mass action law of this equilibrium electronic 'chemical' reaction is given by the following equation, (242.1), which states that the product of the concentration of the reacting species, namely the electrons and holes, is proportional to the product of the concentration of the resulting species which is the perfect lattice in this case. The proportionality constant is known as the **equilibrium constant**, which is not a function of the concentrations and is only a function of temperature. Mathematically, this is written as

$$NP = Kn_V$$

where K is the equilibrium constant and n_V is the concentration of the covalent bonds which is about 10^{23} bonds/cm^3 in Si and other semiconductors.

If we substitute the expressions for electron and hole concentrations from (233.8) and (233.10) into (242.1), we get

$$\begin{aligned} NP &= N_C\exp[-(E_C-E_F)/kT]\cdot N_V\exp[-(E_F-E_V)/kT] \\ &= N_CN_V\exp[-(E_C-E_V)/kT] = N_CN_V\exp[-E_G/kT] && (242.2) \\ &= n_i^2. && (242.2A) \end{aligned}$$

Thus, the NP product is indeed a constant which is independent of N or P but depends only on temperature. This constant, as shown above, is n_i^2.

Using $NP=n_i^2$, n_i from (241.1) and E_I from (241.2), the electron and hole concentrations given by (233.8) and (233.10) can be rewritten as

$$N = N_C\exp[-(E_C-E_F)/kT] \qquad (233.8)$$

$$= n_i\exp[-(E_I-E_F)/kT] \qquad (242.3)$$

and

$$P = N_V\exp[-(E_F-E_V)/kT] \qquad (233.10)$$

$$= n_i\exp[-(E_F-E_I)/kT]. \qquad (242.4)$$

These alternative expressions of N and P in terms of n_i and E_I, given by (242.3) and (242.4), are particularly powerful to analyze and to understand semiconductor device characteristics since they are "reci-symmetrical". More precisely, the normalized concentrations, N/n_i and P/n_i, are exactly reciprocals of each other. We shall make use of these two expressions to compute the equilibrium electrical properties of semiconductors. We shall also use them to derive the corresponding nonequilibrium expressions when there is a current flow. That nonequilibrium extension will lead logically to the definition of the quasi-Fermi levels at any non-equilibrium conditions as an extension of the Fermi level E_F defined only at thermal equilibrium.

(2) The Charge Neutrality Condition

The second relationship needed to compute the electron and hole concentrations at thermal equilibrium in an impure semiconductor is the **charge neutrality condition.** This condition states that if an excess charge appears in a small volume element inside an nearly uniform impure semiconductor, the excess charge will decay to zero very quickly with a decay time constant that is known as the **dielectric relaxation time** of the material. The dielectric relaxation time is the product of the specific capacitance and conductivity of the material. For example, if a Si crystal has a conductivity of σ = 1 mho/cm or 1 Siemen/cm (resistivity = 1 ohm-cm), then using the dielectric constant of Si of ε_{Si} = 11.8 and ε_o = permittivity of free space = $8.854\text{x}10^{-14}$ Farad/cm, the dielectric relaxation time of this 1 ohm-cm Si is

$$t_d = \sigma^{-1}\varepsilon_{Si}\varepsilon_o = 1\text{x}11.8\text{x}8.85\text{x}10^{-14}\ \text{sec} = 1.04\text{x}10^{-12}\ \text{sec}$$

or about 1 picosecond. This shows that a net or imbalance of electrical charge cannot exist in a relatively uniform semiconductor because if it is introduced into the bulk of the semiconductor, the magnitude of the charge will decay to $1/e=0.3678$ of its original value in a time of 10^{-12} second or 1 picosecond. In several picoseconds, the charge would have completely disappeared (to the surface via Gauss Law due to electrostatic repulsion or drain off to the infinite-size earth if the semiconductor is connected to earth). Thus, a uniform semiconductor is electrically neutral. The equation of electrical neutrality is given by

$$\rho = q[P - N + (N_{DD} - N_D) - (N_{AA} - P_A)] = 0. \qquad (242.5)$$

The symbol ρ is used to denote the volume space charge density (Coulomb/cm^3) which is zero because of electrical neutrality. N_D is the concentration of electrons trapped at the donors. (N_{DD}-N_D) is the concentration of the ionized and positively charged donors. P_A is the concentration of holes trapped at the acceptors. N_{AA}-P_A is the ionized or negatively charged acceptor concentration.

In this initial and elementary analysis, we will assume that all of the impurities are ionized, hence, $N_D=0$ and $P_A=0$. Deionization will be discussed in section 252. Thus, the electrical neutrality condition given by (242.5) simplifies to

$$\rho = q[P - N + N_{DD} - N_{AA}] = 0. \qquad (242.5A)$$

The electron and hole concentrations can now be obtained by solving simultaneously (242.2A) from the mass action law and (242.4) from the neutrality condition. Substituting $PN=n_i^2$ or $P=n_i^2/N$ into (242.5A) for a n-type semiconductor, that is a semiconductor whose donor impurity concentration is greater than its acceptor impurity concentration, $N_{DD} > N_{AA}$, then we have the quadratic equation

$$0 = \rho = q[(n_i^2/N) - N + N_{DD} - N_{AA}]$$

or

$$0 = N^2 - (N_{DD} - N_{AA})N - n_i^2. \qquad (242.6)$$

Thus, the electron concentration is given by

$$N = (1/2)[N_{DD} - N_{AA} + \sqrt{(N_{DD} - N_{AA})^2 + 4n_i^2}]. \qquad (242.7)$$

The hole concentration can be obtained using the above for N and $PN=n_i^2$.

As a numerical example, let $N_{DD}=1.1\text{x}10^{15}$ Phosphorus/cm^3 and $N_{AA}=10^{14}$ Boron/cm^3. At T=300K, $n_i \simeq 10^{10}$ carrier/cm^3 in Si. Thus, N_{DD}-$N_{AA}=1\text{x}10^{14}$ $>>$ n_i and n_i in the square root of (242.6) can be dropped. Then,

$$N \simeq N_{DD} - N_{AA} \text{ and } P \simeq n_i^2/(N_{DD}-N_{AA}). \qquad (242.7A)$$

Plugging in the numbers, the electron and hole concentrations are

$$N = 1.1\text{x}10^{15} - 1.0\text{x}10^{14} = 1.0\text{x}10^{15} \quad \text{electron/cm}^3$$

and

$$P = n_i^2/N = 10^{20}/10^{15} = 1.0\text{x}10^5 \qquad \text{hole/cm}^3.$$

The result shows that the electron concentration in this n-type Si is independent of the intrinsic carrier concentration and depends only on the concentration of the dopant impurities. It is equal to the **net impurity concentration** or donor minus acceptor impurity concentrations. Because $N > P$ in this example, this semiconductor is known as the **n-type semiconductor.** Electrons are the **majority carriers** and holes, the **minority carriers**. This is also known as an **extrinsic semiconductor** since the concentration of its main charge carriers or majority carriers depends only on the extrinsic source, the dopant impurities in this case.

The numerical result also shows that the hole concentration is substantially smaller than the electron concentration in an n-type Si, $P<<<N$. In fact, it is also significantly smaller (five orders of magnitudes smaller) than the intrinsic electron concentration, $P<<<n_i$, where n_i is about 10^{10} electrons/cm^3 in Si.

(3) Criterion of Extrinsic Semiconductor

To find the quantitative criterion that makes the approximation (242.7A) valid and a semiconductor extrinsic, a more rigorous analysis must be made by retaining the higher order terms in the Taylor series expansion of the radical in (242.7). Using the Taylor series expansion, (See H.B.Dwight, **Tables of Integrals and Other Mathematical Data**, Fourth edition, MacMillan, 1961; or similar tables.)

$$\sqrt{(1+x)} = 1+ (x/2) - (1\cdot1\cdot x^2/2\cdot4) + (1\cdot1\cdot3\cdot x^3/2\cdot4\cdot6) - \ldots$$

the algebra are given as follows starting from (242.7).

$$N = \tfrac{1}{2}\{(N_{DD}-N_{AA}) + \sqrt{(N_{DD}-N_{AA})^2 + 4n_i^2}\} \qquad (242.7)$$

$$= \tfrac{1}{2}\{(N_{DD}-N_{AA}) + (N_{DD}-N_{AA})[1 + \tfrac{1}{2}(4n_i^2/(N_{DD}-N_{AA})^2 - \ldots\ldots]\} \qquad (242.8A)$$

$$\simeq (N_{DD}-N_{AA}) + n_i^2/(N_{DD}-N_{AA}) \qquad (242.8B)$$

$$\simeq (N_{DD}-N_{AA}). \qquad (242.8C)$$

The Taylor series expansion of the radical is given in (242.8A) where we define $[2n_i/(N_{DD}-N_{AA})]^2=x$ to simplify notation. The higher order terms, x^2, x^3, ..., are not written out in (242.8A) and are dropped in (242.8B). It is further approximated to give the final solution shown in (242.8C) if $n_i^2/(N_{DD}-N_{AA})$ is much smaller than $(N_{DD}-N_{AA})$. For 1% error, the criterion is then $(N_{DD}-N_{AA}) \geq 100n_i^2/(N_{DD}-N_{AA})$ or

$$N_{DD}-N_{AA} \geq 10n_i \qquad (242.9)$$

or

$$N \geq 10n_i. \qquad (242.9A)$$

The semiconductor is known as **extrinsic** when (242.9A) is satisfied, or when (242.9) is satisfied if all donors and acceptors are ionized.

(4) Components of Carrier Concentration are not Additive

We display the intermediate approximation in (242.8B) for a very specific purpose, that is, to demonstrate a common error made by beginners and some textbook authors and teachers. (242.8B) shows that the electron concentration is not the sum of the electrons thermally released by the donors minus those captured by the acceptors, N_{DD}-N_{AA}, and the intrinsic electrons thermally generated by breaking the covalent bonds, n_i, i.e.

$$N \neq (N_{DD}-N_{AA}) + n_i . \tag{242.10}$$

The reason is that electrical neutrality could not be maintained because n_i gives too many thermally generated intrinsic electrons, i.e. the n_i term is in error. This is demonstrated by the following analysis.

$$
\begin{aligned}
\rho/q &= P - N + (N_{DD}-N_{AA}) \\
&= -\, n_i^2/[(N_{DD}-N_{AA})+n_i] - (N_{DD}-N_{AA}) - n_i + (N_{DD}-N_{AA}) && (242.10A) \\
&\approx -\, n_i^2/(N_{AA}-N_{DD}) - n_i && (242.10B) \\
&\approx -\, n_i \neq 0. && (242.10C)
\end{aligned}
$$

The PN product, $PN=n_i^2$, is used for N in (242.10A) which is further simplified in its denominator by the extrinsic approximation N_{DD}-$N_{AA}>10n_i$ derived in (242.9A). (242.10C) shows that in this erroneous analysis, the error comes from the assumption that the thermally generated electron concentration is n_i. The correct answer is given by (242.8B) which shows that electron concentration contains those from the donors minus the acceptors, N_{DD}-N_{AA}, plus just enough thermally generated electrons from bond breaking to electrically neutralize the holes from bond breaking, $n_i^2/(N_{DD}-N_{AA})=P$. Electrical neutrality is maintained and particle conservation is also satisfied as demonstrated as follows using (242.8B).

$$N \approx (N_{DD}-N_{AA}) + n_i^2/(N_{DD}-N_{AA}) \tag{242.8B}$$

and

$$
\begin{aligned}
P = n_i^2/N &\approx n_i^2/[(N_{DD}-N_{AA}) + n_i^2/(N_{DD}-N_{AA})] \\
&\approx n_i^2/(N_{DD}-N_{AA}) .
\end{aligned}
$$

Thus,

$$N \approx (N_{DD}-N_{AA}) + P$$

or

$$\rho/q = P - N + N_{DD} - N_{AA} = 0$$

which gives the correct result. Thus, the key point is that $PN=n_i^2$ so when N increases, P must decrease in a semiconductor at thermal equilibrium.

(5) Summary of Carrier Concentration Equations

A similar analysis can be made if $N_{AA} > N_{DD}$. Then, $P > N$ and this is a p-type semiconductor where hole is the majority carrier and electron is the minority carrier. A summary of the equations for the n-type and p-type semiconductors are given as follows.

n-type Semiconductor

$N > P$

$N_{DD} > N_{AA}$

EXACT RESULT

$N = (1/2)[N_{DD}-N_{AA} + \sqrt{(N_{DD}-N_{AA})^2+4n_i^2}]$ Majority Carriers (242.11)

$P = n_i^2/n$ Minority Carriers (242.12)

APPROXIMATION
(Accurate if $N_{DD}-N_{AA} > 10n_i = 10^{11}\,\text{cm}^{-3}$ for Si at 300K.)

$N \approx N_{DD} - N_{AA}$ (242.11A)

$P \approx n_i^2/(N_{DD}-N_{AA})$ (242.12A)

p-type Semiconductor

$P > N$

$N_{AA} > N_{DD}$

EXACT RESULTS

$P = (1/2)[N_{AA}-N_{DD} + \sqrt{(N_{AA}-N_{DD})^2+4n_i^2}]$ Majority Carriers (242.13)

$N = n_i^2/P$ Minority Carriers (242.14)

APPROXIMATION
(Accurate if $N_{AA}-N_{DD} > 10n_i = 10^{11}\,\text{cm}^{-3}$ for Si at 300K)

$P \approx N_{AA} - N_{DD}$ Majority Carriers (242.13A)

$N \approx n_i^2/(N_{AA} - N_{DD})$ Minority Carriers (242.14A)

243 Temperature Dependences of N, P, and E_F

Using the foregoing formulae, the concentration of electrons and holes, N and P, and the Fermi level relative to the intrinsic Fermi level, E_F-E_I, in n-type Si are computed at various temperatures. They are plotted as a function of temperature in Figs.243.1(a) and (b). The ionized donor impurity concentration, N_{DD}, is the constant parameter for each curve. If there are also some ionized acceptor impurities in the n-type Si, then N_{DD} is to be replaced by the net ionized impurity concentration, $[N_{DD} - N_{AA}]$, which is >0 in n-type. The temperature range, -100C to 500C, covers the practical range -55C to 125C. These concentration and Fermi level curves can also be used for a p-type Si.

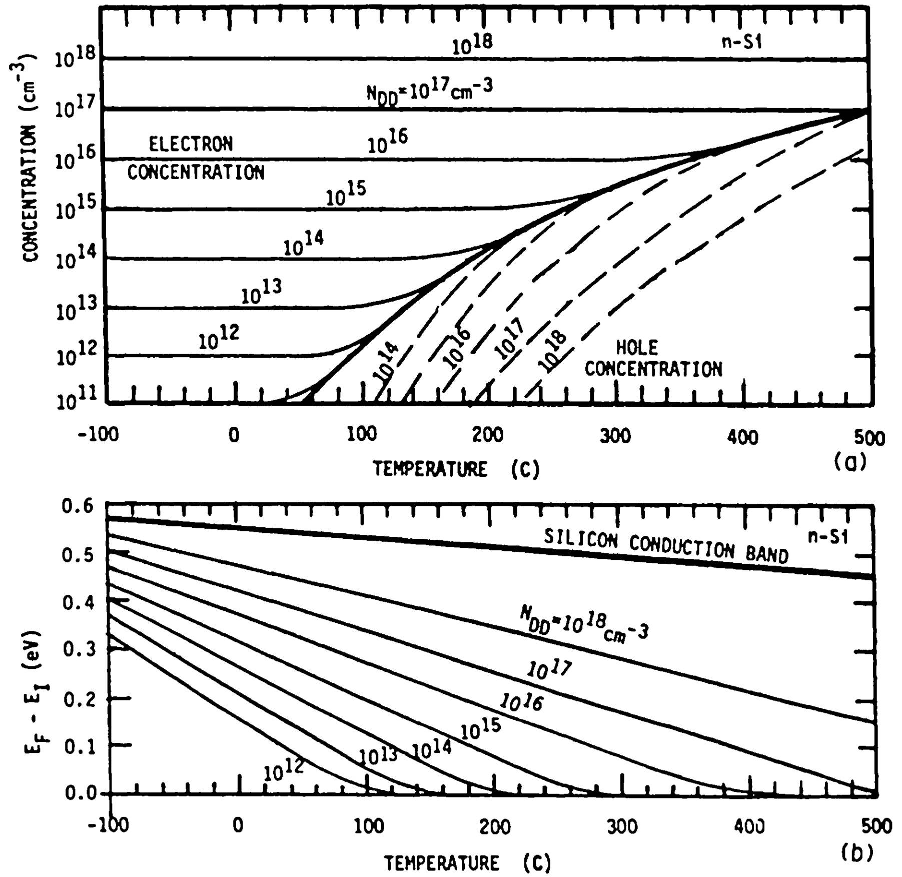


Fig.243.1 (a) The concentrations of electrons and holes and (b) the Fermi level position relative to the intrinsic Fermi level position in Si as a function of temperature.

244 Intrinsic Temperature

Below a certain temperature, Fig.243.1(a) shows that the concentration of electrons (solid curves) and holes (broken curves) vary rapidly with the ionized dopant impurity concentration, N_{DD} in n-Si. The same rapid variations occur also if there are some acceptors, then N_{DD} is replaced by the net ionized impurity concentration, N_{DD}-N_{AA}. The temperature boundary is marked in Fig.243.1(a) by the lowest solid curve above the broken lines. Above this temperature boundary, the dependence on impurity concentration ceases. This is known as the **intrinsic temperature**, T_i. Above T_i, the Fermi level varies more rapidly with N_{DD} as indicated in Fig.243.1(b).

The cause of this independence on impurity concentration above the intrinsic temperature is $n_i >> N_{DD}$ or $n_i >> N_{DD}-N_{AA}$ in the n-type semiconductors. The fundamental reason of large n_i is demonstrated by the transition energy band diagrams of an n-type Si given in Figs.244.1(a) and (b) which contains only N_{DD} donors with energy level E_D, and no acceptors, $N_{AA}=0$.

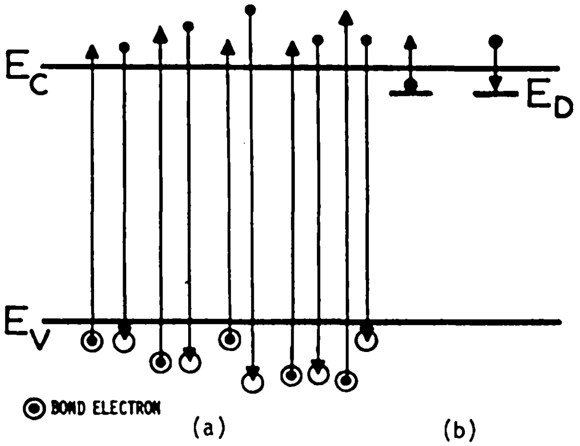


Fig.244.1 The transition energy band diagram in an n-type Si containing N_{DD} donors and no acceptors. (a) Band-to-band electron-hole generation (upward arrows) and recombination (downward arrows) transitions. (b) Band-trap electron emission (upward arrow) and capture (downward arrow) transitions. Circled dot is an electron in a covalent bond which is just about to be broken to create an electron-hole pair.

There are two sources of electrons in the conduction band. Figure 244.1(a) shows the intrinsic source of the electrons from thermal generation of the electron-hole pairs via breaking the covalent or electron-pair bonds by the vibrating host atoms. Figure 244.1(b) shows the extrinsic source of electrons from thermal release or thermal emission of the trapped electron at the donor or trapped hole at the acceptor. **When $T>>T_i$, the intrinsic source (a) gives many more electrons than the extrinsic source (b).**

The inverse of the generation and emission processes are also indicated in Figs.244.1(a) and (b). In Fig.244.1(a), the upward arrows depict **thermal generation of electron-hole pairs** while the downward arrows, **thermal recombination**. To provide vivid illustration, a **circled dot** (a new symbol) is used to denote the electron in the covalent bond that is just about to be broken to create the electron-hole pair. In Fig.244.1(b), the upward arrow depicts the **thermal emission of a trapped electron** while the downward arrow, the **thermal capture of an electron** by the shallow donor trap. At thermal equilibrium, the upward and downward transition rates of the band-to-band transitions in (a) are exactly equal. They are also equal in (b). This balance gives the space-time constant electron, hole and trapped electron concentrations, N, P, and N_D, in a homogeneous semiconductor at thermal equilibrium. In chapter 3, the kinetics of these transitions are described which gives the space-time dependences of the concentrations.

The intrinsic temperature is an important material parameter in semiconductor devices. It is the high temperature limit above which a semiconductor device will lose its useful electrical characteristics. Devices that require a large difference between the electron and hole concentrations can no longer operate properly above T_i because $N=P=n_i$. For example, the p- and n-type layers of a p/n junction diode rectifier will both become intrinsic when $T>T_i$, hence, electrically indistinguishable. Thus, above T_i, the p/n junction no longer exists and its highly nonlinear current-voltage or rectification property is lost. Similarly, above T_i, the p/n junction bipolar transistor will lose its amplification.

Quantitative Definition of T_i

A definition of the intrinsic temperature is derived using n-type semiconductor as the model which has $N>P$ and $N_{DD}>N_{AA}$. From the exact solution of the electron concentration, (242.7), the intrinsic temperature can be defined by the following analysis.

$$N = \tfrac{1}{2}[N_{DD}-N_{AA} + \sqrt{(N_{DD}-N_{AA})^2+4n_i^2}] \tag{244.1}$$

$$= \tfrac{1}{2}[2n_i]\{[(N_{DD}-N_{AA})/2n_i] + 1 + \tfrac{1}{2}[(N_{DD}-N_{AA})/2n_i]^2 + \ldots.\} \tag{244.1A}$$

$$\approx \tfrac{1}{2}[2n_i]\{[(N_{DD}-N_{AA})/2n_i] + 1\} \tag{244.1B}$$

$$= n_i + (N_{DD}-N_{AA}) \tag{244.1C}$$

$$\approx n_i \qquad \text{when } n_i \geq 10(N_{DD}-N_{AA}) \tag{244.1D}$$

and

$$P = n_i^2/N \approx n_i. \tag{244.2}$$

Equation (244.1A) is obtained by Taylor series expansion of the radical in (244.1) about $2n_i$, assuming $4n_i^2 >> (N_{DD}-N_{AA})^2$. Equation (244.1B) is obtained from (244.1A) by dropping the quadratic and higher terms from the Taylor series. Thus, an operational definition of the intrinsic temperature may be obtained from

$$10|N_{DD}-N_{AA}| = n_i = \sqrt{N_C N_V}\exp(-E_G/2kT_i) \tag{244.3}$$

$$= 2.50\times10^{19}(T_i/300)^{3/2}(m_e m_h/m^2)^{3/4}\exp(-E_G/2kT_i). \tag{244.3A}$$

In practice, the transistor characteristics would have deteriorated substantially when n_i has increased to about 10% of N or $N_{DD}-N_{AA}$ in n-type sample.

Equation (244.3) shows that not only high dopant impurity concentration but also large energy gap gives high intrinsic temperatures. In fact, the intrinsic temperature is nearly proportional to the energy gap. Thus, semiconductor devices made in GaAs(E_{G0}=1.52eV) and SiC(E_{G0}>3.0eV) can operate at higher temperatures than in Si(E_{G0}=1.1655eV) before the n-type and p-type regions in the devices become intrinsic. Table 244.1 lists the intrinsic temperature defined by $n_i=10|N_{DD}-N_{AA}|$ rather than $n_i=0.1|N_{DD}-N_{AA}|$ for Si at several impurity concentrations and for Ge, GaAs, GaP, ZnS and SiC at $|N_{DD}-N_{AA}|=10^{15}\text{cm}^{-3}$. Effective masses and energy gaps and their temperature dependences are not known or accurately determined for those labeled with < sign. For these, we assumed $m_e=m_h=m$ and $E_G(T_i) = E_G(T=300K)$ in the calculation which introduce only a small error. The T_i values in Si can be compared with the continuous T_i curve given in Fig.242.1(a). Since T_i of SiC is so high (1300C), the maximum operation temperature of a SiC p/n junction diode is limited by the melting point of the contacting and encapsulating materials. The high T_i and chemical inertness of SiC have attracted continued research interest to develop SiC device technology for high temperature operations.

Table 244.1
Intrinsic Temperatures of Semiconductors

	CONCENTRATIONS		INTRINSIC TEMPERATURE	ENERGY GAP	MELTING POINT
CRYSTAL TYPE	IMPURITY $N_{DD}(cm^{-3})$ Given	INTRINSIC $n_i(cm^{-3})$ Defined	(K and C) Computed	at T_i (eV) Given	(K and C) Given
Si	10^{13}	10^{14}	462K (189C)	1.02	1685K (1412C)
Si	10^{14}	10^{15}	537K (264C)	0.988	
Si	10^{15}	10^{16}	635K (362C)	0.945	
Ge	10^{15}	10^{16}	466K (193C)	~0.6	1231K (957C)
GaAs	10^{15}	10^{16}	909K (636C)	1.19	1510K (1237C)
GaP	10^{15}	10^{16}	<1250K (977C)	<2.24	1750K (1477C)
ZnS	10^{15}	10^{16}	<1917K (1644C)	<3.68	2100K (1800C)
SiC	10^{15}	10^{16}	<1607K (1334C)	<3.0	3070K (2797C)

Components of Carrier Concentration are Additive when $T > T_I$

Equation (244.1C), which is a rearrangement of (244.1B), shows that in the intrinsic temperature range, the electron concentration is indeed the sum of the two sources. It is the sum of the intrinsic carrier concentration n_i, and the electrons thermally released from the donors minus those trapped by the acceptors. This result differs from (242.8B) which showed that the two sources of electrons cannot be algebraically added when the semiconductor is extrinsic or $n_i << N_{DD}-N_{AA}$. The reasons it can be added here are that $n_i >> N_{DD}-N_{AA}$, and electrical neutrality can be maintained by the two sources of electrons at high temperatures: (i) n_i intrinsic electrons and holes are generated in pairs, and (ii) the N_{DD} extrinsic electrons released by the donors are exactly neutralized by the positively charged N_{DD} donors. This is also consistent with the conservation of particle number.

245 Temperature Dependence of the Electron Distribution

A fundamental and application useful question is the dependence of the electron energy distribution on temperature. A first glance at the temperature dependence of the Fermi function shown in Fig.231.2(b) would suggest that there are more electrons at higher energies or higher kinetic energies as temperature is increased. This is the correct answer but the reasoning is incomplete. For example, in the extrinsic temperature range of an n-type semiconductor, the electron concentration is constant, independent of temperature, and equal to the net ionized impurity concentration, $N = N_{DD}$, as indicated by the foregoing analysis and Fig.243.1(a). However, the Fermi energy level, E_F, drops towards the intrinsic Fermi level near the midgap with increasing temperature as indicated in Fig.243.1(b) which would lower the number of high energy electrons. To circumvent these two opposite temperature dependences, the integrand of the electron density integral, (233.7) and (233.7A), is plotted as a function of temperature. From these, we have $dN(E) = n(E,T)dE = f(E,T)D(E)dE = N\cdot(2/\sqrt{\pi})(kT)^{-3/2}\sqrt{E}\exp(-E/kT)dE$ where $N=N_C\exp[-(E_C-E_F)/kT]$ from (233.8) and $f(E)=1/\{1+\exp[(E-E_F)/kT\}\simeq\exp[-(E-E_F)/kT]$. Thus, the density of electrons per unit energy range at energy E is ($E_C=0=$reference)

$$n(E,T) = dN(E)/dE$$

$$= N\cdot(2/\sqrt{\pi})(kT)^{-3/2}\sqrt{E}\exp(-E/kT) \tag{245.1}$$

$$= N_{DD}\cdot(2/\sqrt{\pi})(kT)^{-3/2}\sqrt{E}\exp(-E/kT)\ (\mathrm{cm^{-3}eV^{-1}}) \tag{245.2}$$

where $N=N_{DD}$=constant and independent of E and T. A plot of the normalized electron energy distribution, $n(E,T)/N$ versus E, is given in Fig.245.1 at T=300K, 600K and 1200K. They all have the same area ($\int xdy$) of unity. Comparing the differential area to the left of two of the three curves, it is evident that there are more electrons at higher temperature, i.e. some electrons are moved from lower energies to higher energies as temperature increases. The peak occurs at $E=kT/2$

and the peak value is $n(E_{peak},T)/N = \sqrt{(2/\pi e)}/kT = 0.48394/(kT)$. By integrating (245.1) from $E=E_l$ to $E=\infty$, one can readily prove that the fraction of high energy electron above E_l is given by $N(E>E_l)/N = \mathrm{erfc}(E_l/kT)$ which increases towards 1 with increasing T.

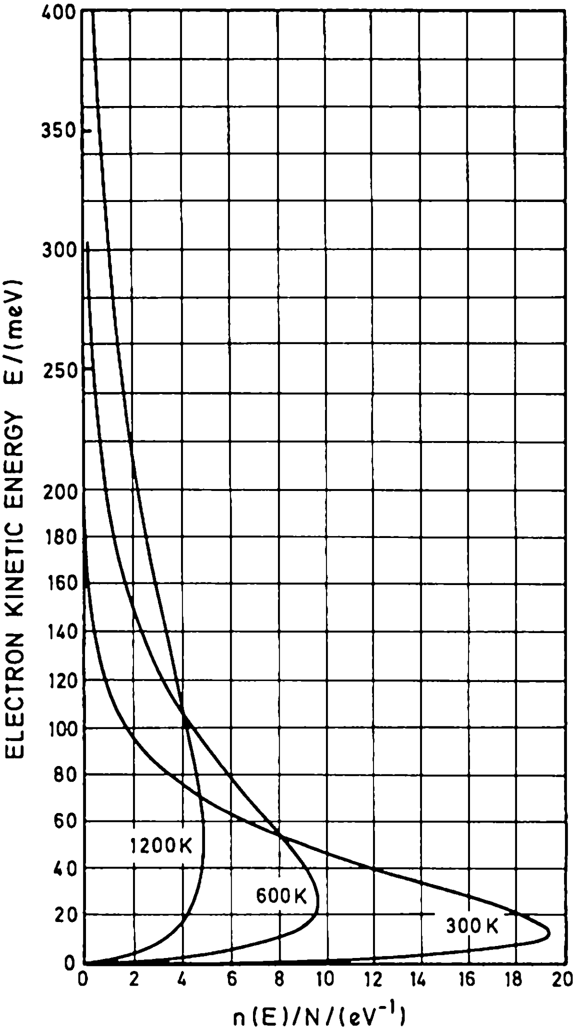


Fig.245.1 The normalized electron energy distribution, n(E)/N, versus energy.

250 DEVICE ESSENTIAL ADVANCED TOPICS

Four advanced equilibrium properties of electronic materials are discussed in the following four subsections, 251-254. They are: Fermi statistics at high carrier concentrations in 251, impurity deionization at high carrier concentration or low temperature in 252, impurity bands in 253, and carrier screening of the impurity bound states in 254. Because of the additional mathematics, they were neglected in the preceding sections in order to crystallize and delineate the basic physics without being obscured by tedious algebra. But they are important semiconductor material properties which must be understood to design and manufacture high performance diodes, transistors, and integrated circuits since device operation frequently involves and often requires high carrier and impurity concentrations. High concentration and low temperature properties are also important in certain photonic devices, such as the far infra-red (10-30 micron) low-temperature intrinsic photon detectors which utilize semiconductors with small energy gap whose electron or hole concentration is quite high at room temperatures even when it is pure. Five specific examples are described below.

The carrier concentration is very high ($>10^{18}\text{cm}^{-3}$) in the emitter and base layers of diodes and bipolar transistors, as well as the heavily doped source and drain regions and the strongly inverted or accumulated surface channels of the MOS transistors. High carrier concentration invalidates the Boltzmann approximation for the Fermi-Dirac distribution function used in (233.6), (233.7A), and (233.10). This high carrier concentration effect is analyzed in section 251.

A significant fraction of the shallow dopant impurities will not be completely ionized at high carrier concentrations due to (i) a high impurity concentration, (ii) a transverse electric field such as the surface space charge layer of a MOS capacitor or MOS transistor, or (iii) a high current density in the emitter or base layer of a bipolar junction transistor operated at high currents. Substantial dopant impurity deionization also occurs at low temperatures even in lowly doped semiconductors, such as the liquid nitrogen temperature, 77.35K, at which future submicron (0.1-micron) Si transistors and integrated circuits may have to operate. This impurity deionization effect is analyzed in section 252.

A high concentration of substitutional impurities on a superlattice will produce an additional periodic potential component which will form impurity bands which alter the energy band and the effective masses. However, the substitutional impurity atoms are seldom located on a superlattice. More often, they are randomly distributed on the periodic lattice. The random distribution generates a spatially random potential energy component which gives an additional perturbation to the energy bands. Furthermore, high impurity concentration will cause the host atoms and the substitutional impurities to displace randomly from the lattice sites. The random displacement to non-lattice sites is equivalent to creating physical defects which further perturb the periodic potential and the energy bands. These are discussed in section 253.

The impurity potential is screened by carriers at high carrier concentrations. The carrier concentration may be so high that the impurity potential can no longer trap an electron or a hole. The condition of disappearance of a bound state is described in section 254.

251 High Carrier Concentration Effects

When the carrier concentration is greater than about one tenth of the effective density of state, $N > 0.1N_C$ in (233.8) or $P > 0.1N_V$ in (233.10), the Boltzmann or exponential approximation of the Fermi function made in (233.6), (233.7A) and (233.10) is no longer valid. The full Fermi function must be used to evaluate the carrier concentration integrals, (233.7) for N and (233.9) for P. The second assumption, the infinite bandwidth approximation $[(E_{C'}-E_C>>kT$ and $E_V-E_{V'}>>kT]$ that made the upper integration limit infinite in (233.7A), is still a good approximation. The high carrier concentration condition is termed degenerate condition, for example, a degenerate semiconductor. This statistical degeneracy or degenerate statistics differs fundamentally from the quantum degeneracies (the spin and spatial or wavefunction configuration degeneracies) described in chapter 1.

From the exact expression given by (233.7), the electron concentration is written in a compact form without making the low concentration Boltzmann approximation to the Fermi function. The algebra is given as follows.

$$N = \int_{E_C}^{E_{C'}} f(E)D(E)dE = \frac{(2m^*)^{3/2}}{2\pi^2\hbar^3}\int_{E_C}^{E_{C'}}\frac{\sqrt{E-E_C}\,dE}{1+\exp[(E-E_F)/kT]} \tag{233.7}$$

$$= \frac{(2m^*)^{3/2}}{2\pi^2\hbar^3}\cdot\frac{\sqrt{\pi}}{2}\cdot\left[\frac{2}{\sqrt{\pi}}\int_0^{\varepsilon_{C'}}\frac{\sqrt{\varepsilon}\,d\varepsilon}{1+\exp[\varepsilon-\varepsilon_F]}\right] \tag{251.1}$$

$$\approx \frac{(2m^*)^{3/2}}{2\pi^2\hbar^3}\cdot\frac{\sqrt{\pi}}{2}\cdot\left[\frac{2}{\sqrt{\pi}}\int_0^{\infty}\frac{\sqrt{\varepsilon}\,d\varepsilon}{1+\exp[\varepsilon-\varepsilon_F]}\right] \tag{251.1A}$$

$$= N_C\cdot F_{1/2}(\varepsilon_F). \tag{251.1B}$$

The integration variable, ε in (251.1) is the normalized electron kinetic energy defined by $\varepsilon=(E-E_C)/kT$, thus, $\varepsilon_F=(E_F-E_C)/kT$. Equation (251.1A) makes the wide conduction bandwidth approximation, $\varepsilon_{c'}=(E_{C'}-E_C)/kT >> 1$.

The compact form of the electron concentration, valid for all (low and high) electron concentrations, is given in (251.1B). N_C was defined by (233.8B). $F_{1/2}(\varepsilon_F)$ is the Fermi-Dirac integral of the one-half order defined by

$$F_{1/2}(\eta) = \frac{2}{\sqrt{\pi}} \int_0^{\infty} \frac{\sqrt{\varepsilon}d\varepsilon}{1 + \exp[\varepsilon-\eta]} . \tag{251.2}$$

The Fermi-Dirac integral cannot be reduced further or represented by an elementary function. Thus, it is numerically integrated and given in tables. Many analytical approximations have been developed since Stoner gave the first approximation in 1936. Blackmore has reviewed the accuracy and range of validity [See J. S. Blackmore, "Approximations for Fermi-Dirac Integrals. Especially the Function $F_{1/2}(\eta)$ Used to Describe Electron Density in a Semiconductor," Solid-State Electronics, 25(11), 1067-1076, November 1982.] Several simple but quite accurate approximations are listed as follows and illustrated in Fig.251.1.

$$F_{1/2}(\eta) \approx \exp(\eta) \qquad -\infty < \eta < -2 \quad \textbf{Boltzmann} \tag{251.2A}$$

$$\approx \exp(\eta)/[1+0.27\exp(\eta)] \qquad -\infty < \eta < 1 \quad \textbf{Blackmore} \tag{251.2B}$$

$$\approx (4/3\sqrt{\pi})[\eta^2 + \pi^2/6]^{3/4} \qquad 1 < \eta < \infty \quad \textbf{Blackmore} \tag{251.2C}$$

$$\approx (4/3\sqrt{\pi})[\eta^{3/2}] \qquad 4 < \eta < \infty \quad \textbf{Sommerfeld} \tag{251.2D}$$

Fig.251.1 indicates that the two Blackmore approximations, (251.2B) and (251.2C), are quite accurate ($\Delta F_{1/2} < 0.3\%$) and indistinguishable from the exact curve which is computed from Nilsson's full range formulae ($\Delta\eta < 0.005$) $\eta = [\log_e u/(1-u^2)] + (3\sqrt{\pi}u/4)^{2/3}/\{1 + [0.24 + 1.08(3\sqrt{\pi}u/4)^{2/3}]^{-2}\}$ where $u = F_{1/2}(\eta)$.

Denormalize the Fermi energy by using $\varepsilon_F = (E_C - E_F)/kT$, then the general electron and hole concentrations valid for all concentrations are given by

$$N = N_C F_{1/2}[(E_F - E_C)/kT] \tag{251.3}$$

and

$$P = N_V F_{1/2}[(E_V - E_F)/kT]. \tag{251.4}$$

The NP product is given by

$$NP = N_C N_V F_{1/2}[(E_F - E_C)/kT] F_{1/2}[(E_V - E_F)/kT]. \tag{251.5}$$

There are two consequences on the NP product due to high carrier concentration. (i) The NP product is smaller than n_i^2, i.e.,

$$NP < N_C N_V \exp[-(E_C - E_V)/kT] = n_i^2. \tag{251.6}$$

This is obvious from Fig.251.1 which shows that the carrier concentration is smaller than that predicted by the Boltzmann approximation, $\exp(\eta)$. (ii) The NP product is no longer independent of the carrier concentrations since it depends on the Fermi energy, E_F, as indicated by (251.5) while the Fermi energy depends on the carrier concentration as indicated by (251.3) and (251.4).

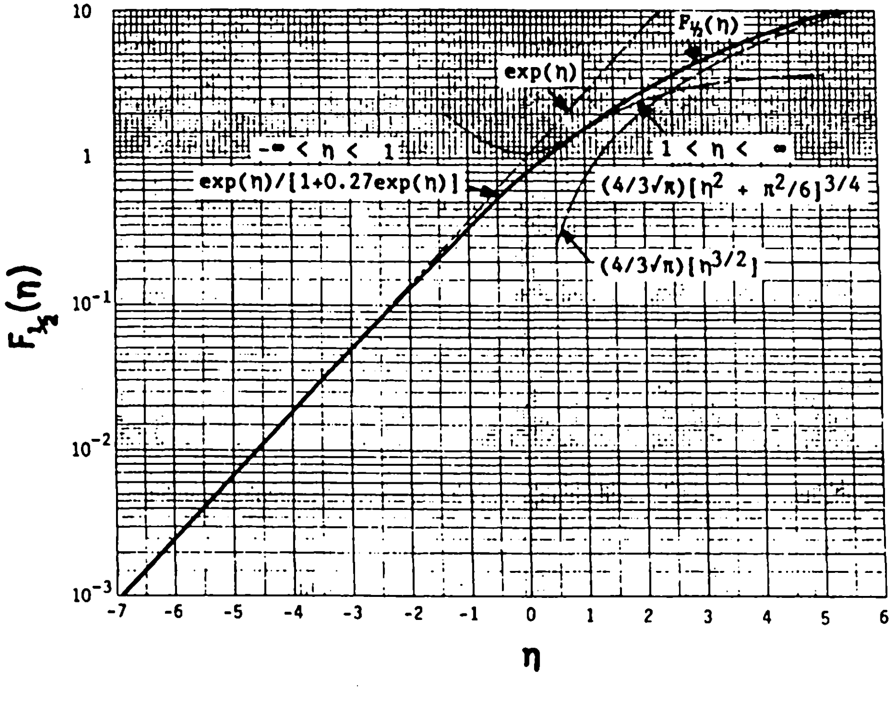


Fig.251.1 The Fermi-Dirac integral of 1/2 order and four of its approximations. The exact curve (solid) is computed from Nilsson's full range formulae but the two Blackmore approximations (joined at $\eta = 1$) are indistinguishable from the exact curve.

A further consequence of (ii) is that the chemical reaction analogy of electron-hole recombination and generation described in section 242(1) must be modified. We used Henry's Law of chemical reaction valid for dilute solutions to arrive at (242.1) which states that the equilibrium reaction constant, K, is a constant independent of the concentrations and dependent only on temperature. This is no longer valid at high electron or hole concentration since the NP product is no longer just a function of temperature but also a function of the concentration of the reacting species, electrons and holes in this case, i.e. $K=K(N,P)\neq$constant. Thus, the chemical analogy used in section 242 must be generalized to (251.5) which is known as the generalized mass action law in semiconductor device physics. One practical example requiring the use of the generalized mass action law is the calculation of N, P, and E_F in small energy-gap semiconductor, such as InSb

($E_G=0.18eV\approx 7kT$ at $T=300K$, $m_e=0.0145m$ and $m_h=0.4m$), where the Boltzmann approximation is no longer valid for electrons because E_F is less than 2kT from the conduction band edge.

252 Impurity Deionization Effects

When we computed the electron and hole concentration in section 242, we used two relationships, (i) the NP product from the mass action law assuming that Henry's law for dilute solution or low carrier concentration is valid, and (ii) the charge neutrality condition assuming that all the donors and acceptors are ionized, that is, no electrons are trapped by the donors and no holes are trapped by the acceptors. In the preceding discussion in section 251, we demonstrated that the mass action law must be generalized. In this section, we show that the assumption of complete impurity ionization must also be removed.

(1) Conditions of Impurity Deionization

There are two conditions under which impurities are not completely ionized, i.e. some electrons are trapped at the donors or some holes are trapped at the acceptors. One of these conditions is low temperature which can be readily understood: at sufficiently low temperatures, $kT << E_C-E_D$ or $kT << E_V-E_A$, thus, there is not enough thermal energy to release the trapped electron from the attractive potential well of the positively charged donor impurity or the trapped hole from the attractive potential well of the negatively charged acceptor impurity.

The second significant impurity deionization condition is high carrier or impurity concentration. Figure 242.1(b) shows that at $N_{DD}=10^{18}cm^{-3}$, the Fermi level is about 3kT below the conduction band edge at T=25C. If the energy level of the electron trapped to the donor impurity is 2kT below the conduction band edge, $E_D=E_C-2kT$, then $E_D-E_F=kT$ which means that an appreciable fraction of the donor impurities is deionized or neutral.

In order to take into account impurity deionization in the charge neutrality condition, (242.5), we need to know the occupation factor of the donor impurities, f_D, and acceptor impurities, f_A, so that the concentration of the neutral donor, N_D, and neutral acceptor, N_A, can be computed from

$$N_D = f_D N_{DD} \tag{252.1}$$

and

$$P_A = f_A N_{AA}. \tag{252.2}$$

This impurity occupation factor is fundamentally different from the Fermi-Dirac distribution function. Unfortunately most textbook authors and teachers have erroneously assumed that they are the same and even called them the Fermi function of the traps. The fundamental difference is that f_D is for the localized electrons trapped to the donors and f_A, the localized holes trapped to the acceptors respectively, while the Fermi function f, (231.6), is for the 'free' or not-bound electrons in the conduction band, and (1-f) is for the 'free' or not-bound holes in the valence band. These not-bound or 'free' electrons and holes are spread out throughout the crystal and not localized to a particular host atom.

The fundamental difference in the occupation of the energy levels by the trap (or bound) and band (or 'free') electrons is that the one-electron bound states (ground and excited states) or the trapping energy levels from an impurity or defect center 'disappear' when an electron is captured or trapped by the center. They disappear when viewed by a second electron. Thus, a second electron cannot be trapped unless the center is a two-electron trap (Like helium or the neutral and negatively charged hydrogen atom described in section 161). However, the electron can be trapped in more than two ways (spin-up and spin-down giving $g_s=2$) by a one-electron center because the wavefunction of the trapped electron may have g_c different spatial configurations of wavefunction orientations and shapes. Thus, the number of ways an electron can be trapped is given by $g_t=g_sg_c$. For example, $g_t=g_sg_c=2x3=6$ at a p-like excited bound state energy level because the electron can be trapped into one of the three p-like spatial configurations, p_x, p_y or p_z. As a second example, the conduction band edge of Si have six equivalent minima located along the six equivalent <100> directions in k-space. Thus, the electron can be trapped at one of the six donor impurity levels from the six band minima, giving $g_c=6$ and $g_D=g_sg_c=2x6=12$ ways for the electron to be trapped to the s-like ground state of a donor impurity Si. After the electron is trapped, the remaining 11 states no longer exist for trapping a second electron. This simple picture of 12-fold degeneracy is modified because the excess impurity core potential over the simple 1/r Coulomb potential due to incomplete core-electron screening will split the $g_c=6$ configuration degeneracy into three energy levels of different spatial configurations and degeneracies with $g_c=1$, 3, and 2. For phosphorus in Si, the three energy levels are at E_D-E_C=-45.105, -33.487, and -32.157meV. Thus, the dominant electron bound state on phosphorus donor in Si has $g_D=g_sg_c=2x1=2$. Since the separations of the three energy levels are small and comparable to kT, the two excited levels cannot be ignored.

(2) Impurity Occupation Factor

The occupation factor of impurity by trapped electron or hole at thermal equilibrium can be derived using the techniques of statistical mechanics to determine the maximum number of ways to place N_D electrons on g_DN_{DD} bound electron states. Obviously, this must include all the excited states. The results are given by

$$f_D = 1/\{1 + (1/g_D)\exp[(E_D-E_F)/kT]\} \qquad (252.3)$$

and

$$f_A = 1/\{1 + (1/g_A)\exp[(E_F-E_A)/kT]\} \qquad (252.4)$$

which can be compared with the Fermi function of the conduction band electrons and valence band holes,

$$f = 1/\{1 + (1/g)\exp[(E - E_F)/kT + \log_e(g)]\}$$

$$= 1/\{1 + \exp[(E - E_F)/kT]\} \quad \text{(electrons)} \qquad (252.5)$$

and

$$f_h = 1 - f$$

$$= 1/\{1 + \exp[(E_F - E)/kT]\} \quad \text{(holes).} \qquad (252.6)$$

The similarity between (252.3) and (252.5) for band and bound electrons, and between (252.4) and (252.6) for band and bound holes is evident. f_D has exactly the same form as the Fermi function if the degeneracy factor, g_D, is combined with E_D by defining an effective donor energy, $E_{D'} = E_D - kT\cdot\log_e(g_D)$ first noted by Shockley and Read, as demonstrated by the following algebra. This is the source of confusion among textbook authors, teachers and device engineers.

$$f_D = 1/\{1 + (1/g_D)\exp[(E_D - E_F)/kT]\} \qquad (252.3)$$

$$f_D = 1/\{1 + \exp[(E_D - kT\cdot\log_e g_D - E_F)/kT]\} \qquad (252.7A)$$

$$f_D = 1/\{1 + \exp[(E_{D'} - E_F)/kT]\} \qquad (252.7B)$$

$$\equiv f(E_{D'}) \qquad (252.7C)$$

(3) Impurity Deionization Examples

We shall now illustrate the reduction of the electron concentration due to impurity deionization in an n-type semiconductor. Two examples are given, the high impurity concentration and the low temperature dependences. There are two ways to solve the problem mathematically, (i) the direct or brute-force approach using the formulae given above and the charge neutrality condition, and (ii) the mass action law approach which simplifies the algebra.

Deionization at High Impurity Concentration

Consider first the high impurity concentration example of a n-type semiconductor doped with N_{DD} donors with a energy level E_D. The charge neutrality condition is then

$$0 = P - N + N_{DD} - N_D \qquad (252.8)$$

$$\approx \quad - N + N_{DD} - N_D \qquad (252.8A)$$

since $P << N$ at all temperatures and donor impurity concentrations. Using f_D of (252.3) in $N_D = f_D N_{DD}$ of (252.1), and the general expression for the electron concentration, then (252.8A) becomes

$$N = N_{DD} - N_D = N_{DD}(1 - f_D) \tag{252.9}$$

$$= N_{DD}/\{1 + g_D\exp[(E_F-E_D)/kT]\} \tag{252.9A}$$

$$= N_C F_{1/2}[(E_F-E_C)/kT]. \tag{252.9B}$$

The last two expressions give

$$N_{DD}/N_C = F_{1/2}[(E_F-E_C)/kT]\cdot\{1+g_D\exp[(E_F-E_C+E_C-E_D)/kT]\}. \tag{252.10}$$

It can be solved to give E_F as a function of N_{DD} at a given temperature. This E_F is then used in (252.9B) to give N, and in (252.10) or (252.3) to give f_D. (252.9A) also shows that the electron concentration is smaller than N_{DD} by the ratio

$$N/N_{DD} = 1/\{1 + g_D\exp[(E_F-E_D)/kT]\} \tag{252.11}$$

because some of the electrons are trapped at the donor.

The effects of impurity deionization are illustrated in Figs.252.1(a) and (b) for an n-type semiconductor doped by N_{DD} donor impurity with $g_D=2$ and $(E_C-E_D)/kT=2$. This corresponds to the phosphorus donor ground state in n-Si at 261.7K (-11.4C). Four curves for each group are plotted to show the effects of impurity deionization as well as Fermi statistics. The four curves correspond to the following four combinations of assumptions:

B,I	Boltzmann Statistics and Impurity Ionization
F,I	Fermi Statistics and Impurity Ionization
B,D	Boltzmann Statistics and Impurity Deionization
F,D	Fermi Statistics and Impurity Deionization

The normalized donor impurity concentration, N_{DD}/N_C, is the independent variable along the abscissa (x-axis). The following parameters are plotted in figure (a): the normalized electron concentrations, N/N_C; the fraction neutral donors or donors occupied by electron, $f_D=N_D/N_{DD}$; and the fraction of ionized donors, $1-f_D=1-(N_D/N_{DD})$.

Consider the donor concentration $N_{DD}/N_C=1$ which is approximately $N_{DD}=N_C\approx2.24\times10^{19}\text{cm}^{-3}$ in Si as computed from (233.8C) by using T=261.7K and $m_e=1.065m$. Figure 252.1(a) shows that Fermi statistics has no effect if all the impurities are assumed ionized and only a very small effect when impurity deionization is included. However, deionization has a tremendous effect, reducing the electron concentration by a factor of 4, from N_{DD} to $0.23N_{DD}$ when deionization is included. The reduction is 0.54 at 10-times lower donor concentration, $N_{DD}/N_C=0.1$ or $2.24\times10^{18}\text{cm}^{-3}$, and 0.85 at 100-times lower donor concentration, $N_{DD}/N_C=0.01$ or $2.24\times10^{17}\text{cm}^{-3}$. If Boltzmann statistics is used with deionization, the reduction is smaller or underestimated but the underestimate is less than 2% at $N_{DD}/N_C=1$ as indicated by comparing the exact (F,D) and approximate (B,D) N/N_C curves shown in Fig.252.1(a).

The fraction of occupied or neutral donor, f_D, and unoccupied or ionized donor, $1\text{-}f_D$, are also plotted as heavy curves in Fig.252.1(a). The Boltzmann approximation for the fraction of ionized donors, $1\text{-}f_D$, is given by the light curve, showing again Boltzmann statistics gives only a small overestimate of $1\text{-}f_D$ at $N_{DD}/N_C=1$.

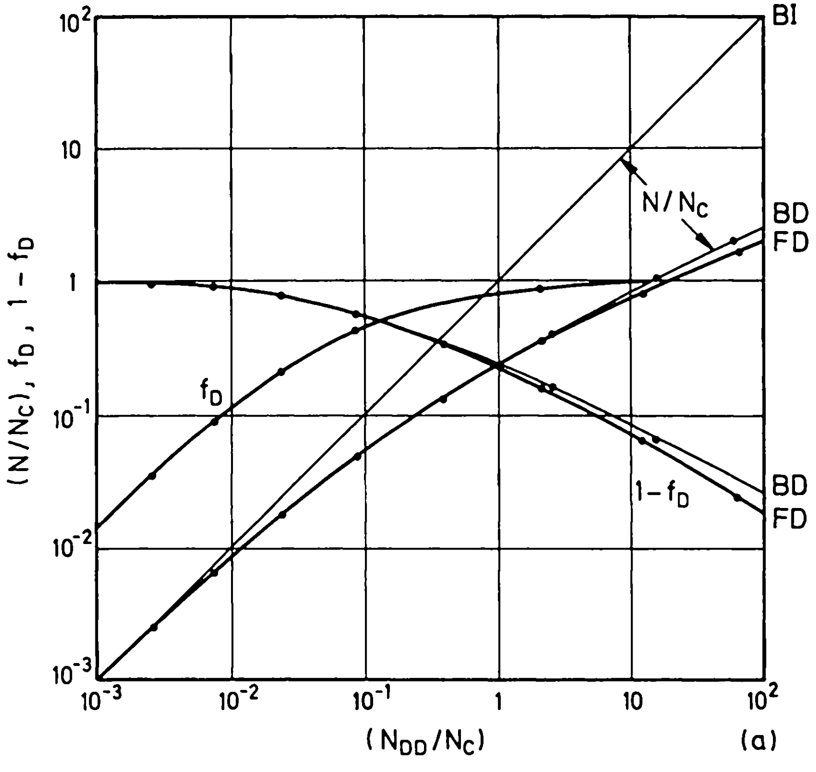


Fig.252.1 The effects of impurity deionization and Fermi statistics on the properties of n-type semiconductor as a function of the total donor concentration, N_{DD}, with $g_D=2$ and $(E_C-E_D)/kT=2$, corresponding to phosphorus in n-Si at T=261.7K (-11.4C) and $N_C=2.24\times10^{19}\text{cm}^{-3}$. Four curves from four assumptions are shown for each group of curves: (B,I), (F,I), (B,D) and (F,D) where B=Boltzmann approximation, F=Fermi statistics, I=Impurity completely ionized, and D=Impurity Deionization included. (a) N/N_C, $f_D=N_D/N_{DD}$, and $1\text{-}f_D$. (b) $(E_F-E_C)/kT$.

In Fig.252.1(b), the Fermi energy level position is plotted as a function of the total donor impurity concentration for the four assumptions. Again deionization has the largest effect which is evident when comparing the (F,I) and (F,D) curves or (B,I) and (B,D) curves. However, when deionization is included as indicated by the (B,D) and (F,D) curves, the Boltzmann approximation gives a very small error even at $N_{DD}/N_C = 10$ or $N_{DD} \approx 2.24\times10^{20}\text{cm}^{-3}$. But, this is deceptive since in semiconductor measurements and device applications, the carrier concentration is the important parameter and a small error in E_F produces a large error in N because $N \propto \exp(E_F/kT)$ at low N and $N \propto E_F^{3/2}$ at high N.

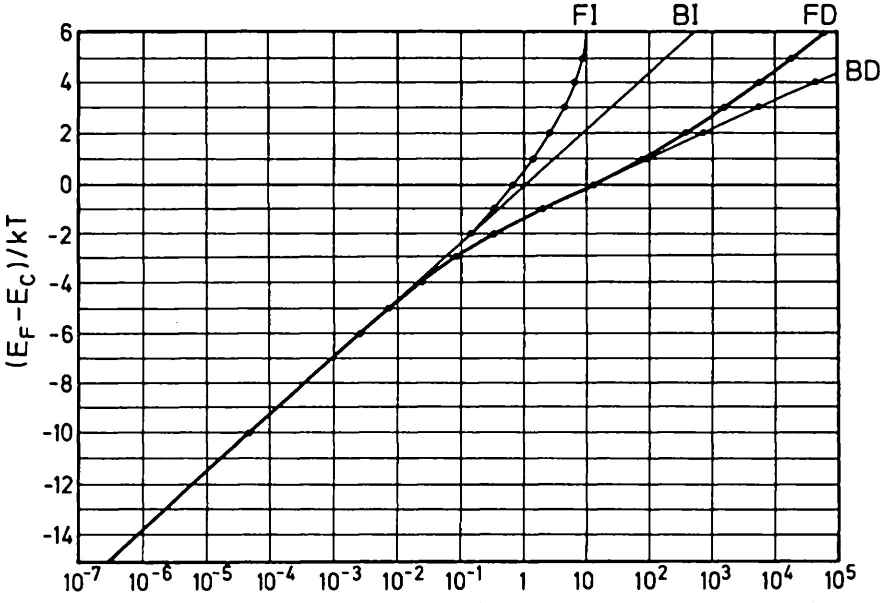


Fig.252.1 The effects of impurity deionization and Fermi statistics on the properties of n-type semiconductor as a function of the total donor concentration, N_{DD}, with $g_D = 2$ and $(E_C - E_D)/kT = 2$, corresponding to phosphorus in n-Si at $T = 261.7$K (-11.4C) and $N_C = 2.24\times10^{19}\text{cm}^{-3}$. Four curves from four assumptions are shown for each group of curves: (B,I), (F,I), (B,D) and (F,D) where B = Boltzmann approximation, F = Fermi statistics, I = Impurity completely ionized, and D = Impurity Deionization included. (a) N/N_C, $f_D = N_D/N_{DD}$, and $1-f_D$. (b) $(E_F - E_C)/kT$.

The isolated impurity bound state model is no longer valid at high impurity concentrations when the inter-impurity distance is comparable to the Bohr radius.

The trapped electron can no longer be localized or trapped at the impurity, and an impurity band is formed. This is discussed in the following section, 253.

Deionization at Low Temperatures

The results of high impurity concentration just described show that impurity deionization has a very large effect on the carrier concentration at room temperature. If impurity deionization is not included, the computed carrier concentration at a given total impurity concentration of N_{DD} is too high. At $N_{DD}=0.1N_C \approx 2.24\text{x}10^{18}\text{cm}^{-3}$, the full-ionization value of N (i.e. $\equiv N_{DD}$) is $1/0.55 = 1.82$ or 82% too high. Impurity deionization is even more important at low temperatures because the thermal vibration energy of the lattice or host atoms, kT, is lower so that the trapped electron is less likely to be thermally emitted or released from the donor impurity.

The charge neutrality equation and the derived equations at high impurity concentrations just discussed, (252.8) to (252.11), are all applicable at low temperatures. Consider first a lowly doped n-Si in which the Boltzmann approximation to the Fermi function is valid for computing the electron concentration. Then, (2.52.9B), (252.10) and (252.11) simplify to

$$N/N_C = F_{\frac{1}{2}}[(E_F-E_C)/kT] \tag{252.9B}$$

$$\approx \exp[(E_F-E_C)/kT] \tag{252.12}$$

$$N_{DD}/N_C \approx F_{\frac{1}{2}}[(E_F-E_C)/kT]\cdot\{1+g_D\exp[(E_F-E_C+E_C-E_D)/kT]\} \tag{252.10}$$

$$\approx \exp[(E_F-E_C)/kT]\cdot\{1+g_D\exp[(E_F-E_C+E_C-E_D)/kT]\} \tag{252.13}$$

$$\equiv (N/N_C)\cdot\{1 + N/K_D\} \tag{252.14}$$

$$N/N_{DD} \equiv 1/\{1 + g_D\exp[(E_F-E_D)/kT]\} \tag{252.11}$$

$$\equiv 1/\{1 + N/K_D\} \tag{252.15}$$

where

$$K_D \equiv (1/g_D)N_C\exp[-(E_C-E_D)/kT] \tag{252.16}$$

is the **equilibrium reaction constant** for the electron trapping-detrapping reaction. This reaction constant can be derived using the transition energy band diagram Fig.244.1(b) for the electron-donor transitions and the following analogous chemical reaction equations

$$\text{Electron + Ionized Donor} \longleftrightarrow \text{Neutral Donor} \tag{252.17}$$

or

$$N \;+\; (N_{DD}\text{-}N_D) \;\longleftrightarrow\; N_D. \tag{252.18}$$

The equilibrium constant can be directly obtained from (252.18) using Henry's Law for dilute solutions, in this case, the low electron concentration so that Boltzmann's approximation is valid for N. Henry's law then states that $N(N_{DD}-N_D)=K_DN_D$ where K_D is independent of concentration. This gives $K_D=N(N_{DD}-N_D)/N_D$ =(252.16) when $f_D=N_D/N_{DD}=1/\{1+(1/g_D)\exp[(E_D-E_F)/kT]\}$ from (252.3) is used. Expressing the equations in terms of K_D simplifies the algebra substantially in the numerical solution as we shall notice in the following discussions.

From the above equations, two temperature ranges can be delineated in addition to the high temperature intrinsic range which is not covered by the above equations since it was assumed that $P<<n_i<<N$ and P was dropped from the space charge density. The two temperature ranges are: the **extrinsic range** (also known as the saturation range) and the **deionization range** (also known as the freeze-out range). The extrinsic range covers the room temperature and the deionization range begins at a lower temperature. In the extrinsic range, essentially all the impurity atoms are ionized. Thus, the exponential term in the denominator of (252.11) or (252.14) can be dropped compared with 1 and

$$N \simeq N_{DD}. \qquad \text{(Extrinsic Temperature Range)} \qquad (252.19)$$

At very low temperatures, a substantial fraction of the donors are deionized or neutral and occupied an electron. And the exponential term in the { } of (252.13) is much larger than 1, or $N/K_D>>1$ in (252.14) so 1 can be dropped:

$$N_{DD}/N_C \simeq (N/N_C)\cdot(N/K_D). \qquad (252.20)$$

Thus, the electron concentration at very low temperatures is given by

$$N = \sqrt{N_{DD}K_D} \qquad (252.21)$$

$$= \sqrt{(N_{DD}N_C/g_D)}\cdot\exp[-(E_C-E_D)/2kT]. \quad \text{(Deionization Range)} \quad (252.21A)$$

Figure 252.2(a) shows a family of $\log_{10}N$ versus 1000/T curves, for an n-Si doped by $N_{DD}=10^{15}$ to 10^{18} phosphorus/cm^3 with $g_D=2$ and $E_C-E_D=0.046$eV. This semi-log plot is used to delineate the three straight line ranges since their slopes give the thermal activation energies which can be used to compute the basic properties of the material from experimental data. The three ranges are: (i) the intrinsic range where $N=P=n_i=\sqrt{(N_CN_V)}\exp[-(E_G/2kT)]$ so the slope is $-E_G/2k$ and gives half of the energy gap, (ii) the extrinsic range where $N=N_{DD}>>P$ so that a measurement of N gives the donor impurity concentration, N_{DD}; and (iii) the deionization range, given by (252.21A) which shows that the slope is $(E_C-E_D)/2k$ and gives the half of the electron binding energy at the donor.

Figure 252.2(b) plots $1-f_D$ $[=N/N_{DD}=(N_{DD}-N_D)/N_{DD}]$ and N versus N_{DD} with the temperature as the parameter. They clearly show increasing deionization or lower N as the temperature is lowered. For example, at the liquid nitrogen temperature, 77.35K, $N=0.73\times10^{15}$cm^{-3} for $N_{DD}=1.0\times10^{15}$cm^{-3} or 27% of the donors are deionized or neutral, each with a trapped electron. At $N_{DD}=10^{18}cm^{-3}$, $N=4.3x10^{16}cm^{-3}$ or nearly 96% of the donors are deionized.

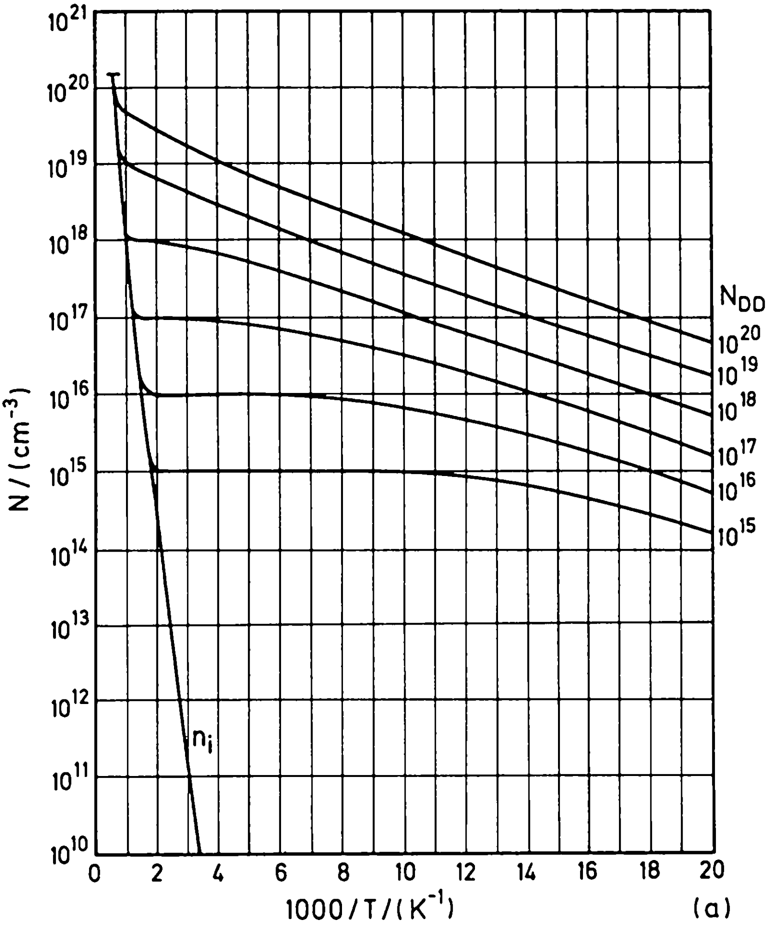


Fig.252.2 Electron concentration in phosphorus doped n-Si with E_C-E_D=0.0455eV and g_D=2. (a) Electron concentration, N, versus 1000/T(K) with N_{DD}=constant parameter. (b) Fraction of donor ionized, N/N_{DD}=1-f_D, and electron concentration, N, versus donor concentration, N_{DD}, with T=constant parameter. (b) on following page.

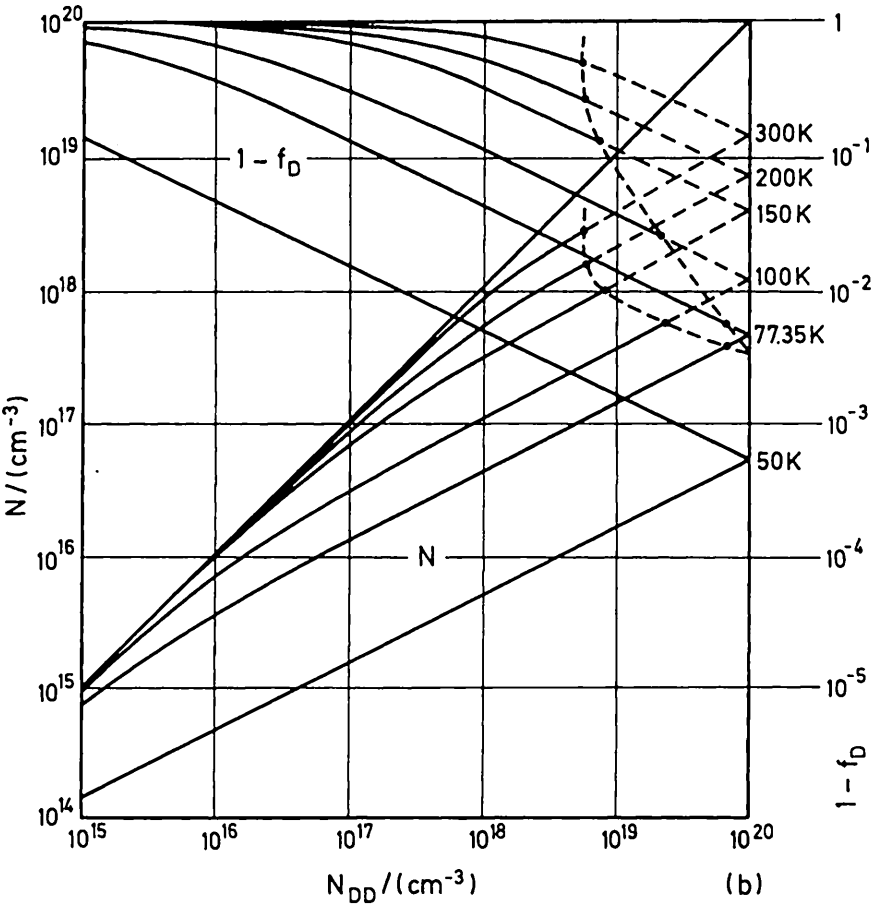


Fig.252.2 Electron concentration in phosphorus doped n-Si with E_C-E_D=0.0455eV and g_D=2. (a) Electron concentration, N, versus 1000/T(K) with N_{DD}=constant parameter. (b) Fraction of donor ionized, N/N_{DD}=1-f_D, and electron concentration, N, versus donor concentration, N_{DD}, with T=constant parameter. (a) on previous page.

The points at which $N/N_C = 1.0$ are also shown which give the boundary (in heavy broken lines) above which Fermi statistics must be used. In the previous example, Fig.252.1(a), we have shown that the Boltzmann approximation will make only a slight error, giving an electron concentration less than 2% higher at $N_{DD}/N_C = 1.0$ than the exact result using Fermi statistics. Thus, the results using Boltzmann approximation given in Figs.252.2(a) and (b) are accurate even at

$N_{DD} = 10^{20}\text{cm}^{-3}$. One can use Fermi statistics for the higher concentration range but an iterative computer calculation is then needed instead of using a hand-held calculator with which these curves are computed.

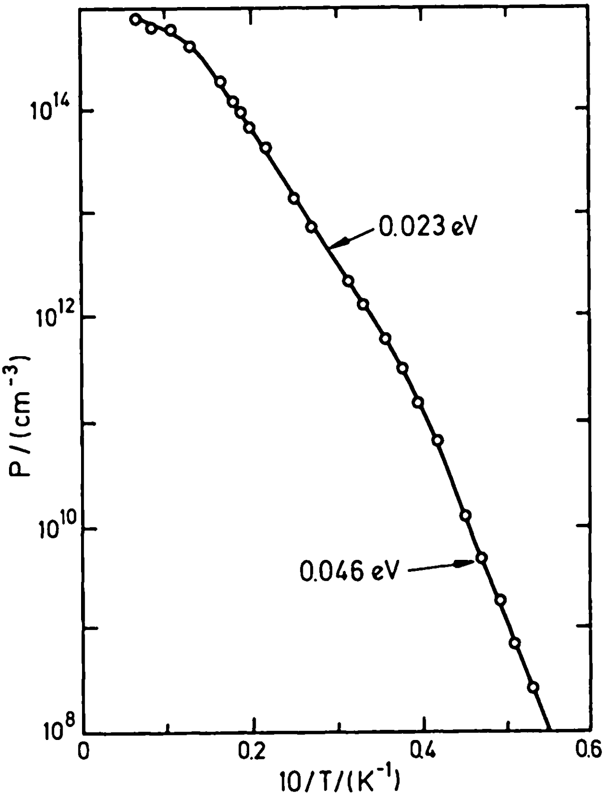


Fig.252.3 Hole concentration versus 1000/T of a p-type Si. Solid line is the theory using $E_A - E_V = 0.046\text{eV}$, $N_{AA} = 7.4\times10^{14}$ boron/cm^3 and $N_{DD} = 1.0\times10^{11}$ phosphorus/cm^3. (From N.B.Hannay, **Semiconductors**, p.31, Reinhold, 1959.)

An experimental data of a p-Si is shown in Fig.252.3 which contains $N_{AA} = 7.4\times10^{14}$ boron/cm^3 and $N_{DD} = 1.1\times10^{11}$ phosphorus/cm^3. The agreement with the theory just presented is excellent for $T \geq 25\text{K}$ or $(10/T) < 0.40$. For $T < 25\text{K}$, a third straight line appears which comes from a minute amount of donor which pins the Fermi level to the acceptor energy, $E_F \to E_A$, when $P < N_{DD}$. This gives an activation equal to $E_A - E_V$ instead of $(E_A - E_V)/2$ at the higher temperatures as indicated by (252.21A). A more exact formulae with $N_{AA} \neq 0$ is given in the problem set.

253 Impurity Bands

When the donor impurity concentration is very high, the bound electron wavefunction on adjacent donors overlaps. The trapped electron is no longer bound or localized to a particular donor impurity center. It is delocalized and can be found on the adjacent impurity atoms because it tunnels through the impurity potential hill separating adjacent impurity atoms. An impurity band is formed. The impurity bandwidth is proportional to the overlap of the adjacent bound electron wavefunctions. This was the qualitative model used in section 182 to estimate the conduction and valence bandwidths in pure crystals.

Experimental data of electron concentration versus temperature, such as the data of Si crystals #126 and #140 in Fig.253.1, showed that the electron concentration is higher and the temperature dependence is weaker than the non-impurity banding theory given in Fig.252.2(a). At the highest impurity concentration in this figure, about 10^{20}cm^{-3} for Si crystal #140, nearly metallic behavior (weak or no temperature dependence) of the electron concentration was observed. These data suggested that the electrons can no longer be trapped at the donor impurity centers and the donor bound states have disappeared.

Impurity band formation is one school of theory to explain the disappearance of impurity bound states. Another theory is based on the presumption that the donor impurity potential is screened by the high concentration of electrons. However, this screening theory has a fundamental flaw when applied to high donor impurity concentrations since on the average, there is only one electron around one donor atom because the electron comes from the donor atom. It is a valid theory only when there are other sources of electrons (or holes), making the electron concentration much higher than the impurity concentration so that there are many electrons (or holes) to screen each impurity atom. Such a situation does exist in the strongly inverted or accumulated surface channel of a MOS capacitor and MOS transistor, and in the base layer of a p/n junction diode or the base and collector layers of a bipolar transistor at very high current densities or under intense illumination. Impurity bound state disappearance due to carrier screening is described in the next section, 254. In this section, we shall estimate the critical impurity concentrations for the disappearance of the impurity bound state due to impurity band formation.

We shall describe two simple models to calculate the donor impurity concentration at which the electron bound state at the donor disappears due to overlap of the bound electron wavefunction on adjacent impurity centers. This impurity concentration is known as the **semiconductive-to-metallic transition concentration**, **the critical concentration**, or the **delocalization concentration**.

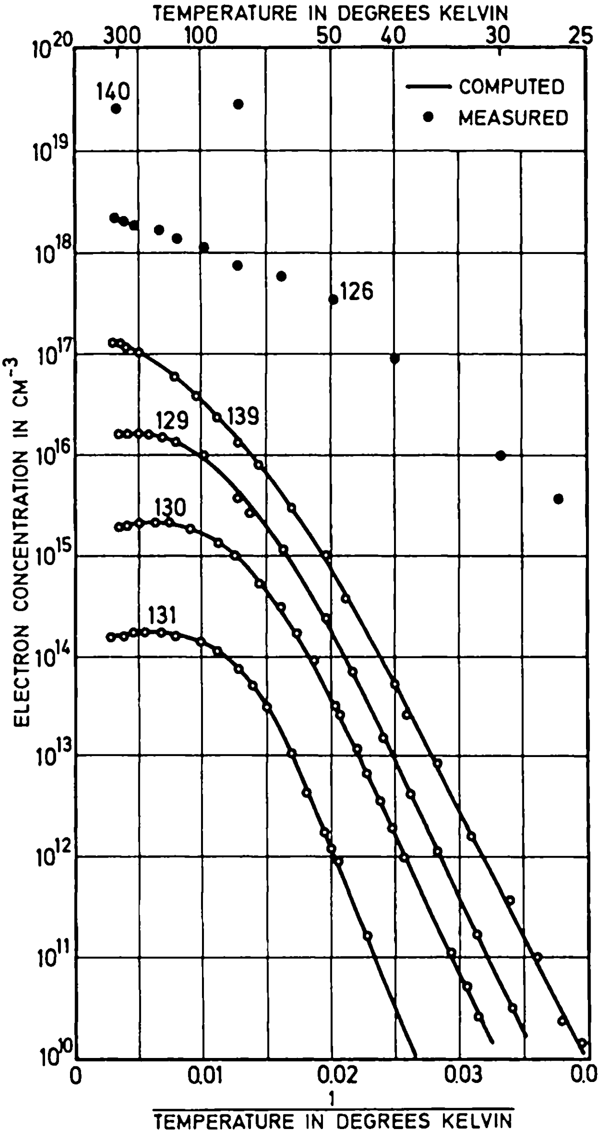


Fig.253.1 Electron concentration versus reciprocal temperature for phosphorus doped n-type Si with $N_{DD} = 2\times10^{14}$ to 10^{20}cm^{-3}. (From Morin and Maita, Phys. Rev. 96,28,1954.)

The first model is the traditional model using the Bohr radius which can give only a hand wavering argument and which gives too high an impurity concentration. The second is a new model devised by this author, which is conceptionally simpler and quantitatively more precise, and which focuses on the height of the impurity potential hill that separates adjacent impurity atoms. Both model scales with the Bohr radius of the electron bound to the impurity atom, a conceptually simple and desirable feature that allow scaling to other impurities once their experimental binding energies are determined.

In the first model (the Bohr radius model), the Bohr radius, a_n given in sections 223 and (141.6), is taken as the critical impurity separation,

$$s_{crn} = 2a_n. \tag{253.1}$$

Then, the critical impurity concentration is

$$N_{cr} = s_{crn}^{-3} = 1/(8a_n^3) \tag{253.2}$$

above which an impurity band is formed and the electron can no longer be localized at the impurity. Thus, in this model, the impurity bands will be formed for the highest excited states, $n = \text{large} \to \infty$ and $a_n = \text{large} \to \infty$, even at very low impurity concentrations. Thus, the band edge is ill-defined and is spatially modulated by the impurity potentials. This band edge potential modulation will alter the curvature of E-k diagram and hence the effective mass of the electron or hole. Since the deionization effect starts by trapping the electron to the lowest donor bound state, $n=1$, deionization will no longer be possible when the ground state disappears or the ground state impurity band is formed. Using the numbers taken for the phosphorus donor in Si given in section 223, $E_1 = 44\text{meV}$, $m^* = 0.45m$ and $a_1 = 13.8\text{A}$, then the critical phosphorus donor concentration is $N_{cr}(a_1) = (2a_1)^{-3} = (13.8\times10^{-8})^{-3} = 3.8\times10^{20}\text{cm}^{-3}$. This is considerably higher than the experimental N_{DD} indicated in Fig.253.1, about $5\times10^{18}\text{cm}^{-3}$, at which deionization seems to begin to disappear.

The second and better estimate focuses on the height of the impurity potential hill that separates the adjacent impurity atoms. This is illustrated in Fig.253.2 for two donor impurities separated by $8r_1$. The figure shows that the bound state solution E_1 disappears when the impurity atoms have a interatomic separation of s_{cr} ($>8r_1$, to be derived) because the impurity potential maximum is equal to the ground state energy, E_1, therefore there is no potential hill separating the two donors. Since second neighbor and distant impurities will further lower the impurity potential maxima, the figure shows the minimum separation. We estimate s_{cr} numerically by considering only two adjacent donor impurity atoms. The electron potential energy due to one positively charged donor impurity is $V(r) = -q^2/(4\pi\varepsilon_s r)$ where ε_s is the dielectric constant of Si. Thus, for two adjacent donors separated by a distance of $2x$, the potential peak has a value of $2V(x)$. The critical separation $s_{cr} = 2x_{cr}$ is when

$$2V(r) = -2q^2/(4\pi\varepsilon_s x) = E_1 = -q^2/[2(4\pi\varepsilon_s a_1)] \qquad (253.3)$$

giving $x_{rc}=4a_1$. The critical separation and impurity concentration are then

$$s_{cr} = 8a_1 \qquad (253.4)$$

and

$$N_{cr} = 1/(512a_1{}^3). \qquad (253.5)$$

This is $4^3=64$ smaller than the $n=1$ Bohr radius model given by (253.2). Using again the numerical values adopted in section 223, $a_1=(4\pi\varepsilon_s)\hbar^2/m^*q^2=13.8A$ where $\varepsilon_s=11.8x8.854x10^{-14}$ F/cm and $m^*=0.45m$, then

$$s_{cr} = 8a_1 = 8x13.8A = 110A \qquad (253.6A)$$

and

$$N_{sc} = 7.43x10^{17}\ cm^{-3} \qquad (253.6B)$$

which is in better agreement with the experimental observations.

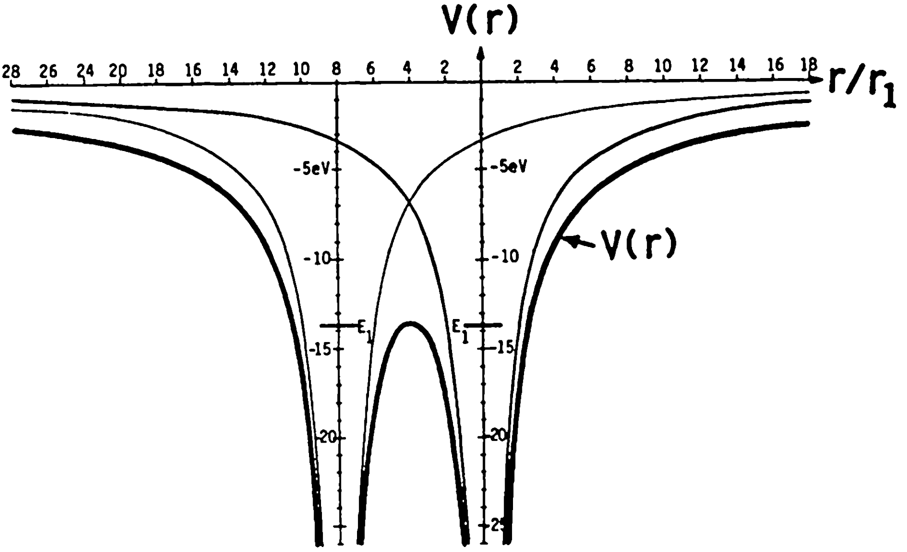


Fig.253.2 The impurity potential hill model for estimating the critical impurity concentration at which the electron bound state disappears.

In addition to the two-impurity assumption and the inherent approximation in the hydrogen point-charge model, several additional factors can account for the graduate decrease (instead of abrupt disappearance) of the experimental activation energy and the higher experimental electron concentration than the theoretical prediction. These are: (1) conduction by electrons in the impurity band which has higher mass and lower mobility, making N higher than the theory from N_{DD}-N_D; (ii) random spatial distribution of impurities so that some impurity atoms are sufficiently far apart ($>s_{cr}$) to still have the isolated bound states, E_n, which would lower N, and some impurity atoms are so close that an electron ground state deeper than E_I may appear since the impurity pair is like a +2q charge, resulting in a distribution of electron binding energy; and (iii) random potentials from random host atom displacements caused by the random spatial distribution of the impurities, which random host potentials may create new bound states for majority as well as minority carriers which would lower N. The random spatial distribution of the impurity atoms is probably the most important factor.

254 Carrier Screening of Impurity

When the electron density is much higher than the donor impurity density, each impurity can then be electrically screened by many electrons which weaken the donor impurity potential to bind or trap an electron. When the electron concentration is sufficiently high, there is no bound state solution of the Schrödinger equation. This gives an estimate of the critical electron concentration at which the impurity activation energy drops to zero.

This carrier screening is known as **Debye-Huckle screening** in dilute electrolytes and semiconductors with low electron concentration. It is known as the **Fermi-Thomas screening** in degenerate semiconductors and metals. The 1/r Coulomb impurity potential is screened and reduced by an exponential function, $\exp(-k_s r)$ where $k_s^{-1} = r_s$ is known as the screening length. It is the distance at which the potential drops by 1/e. The screened Coulomb potential energy of an electron is then

$$V(r) = -\frac{q^2}{[4\pi\epsilon_s r]}\exp(-r/r_s). \qquad (254.1)$$

Note there are two screening sources: (i) the valence electrons denoted by the dielectric constant, ϵ_s, known as **dielectric screening** and (ii) the carriers or conduction band electrons and valence band holes, denoted by r_s. Fig.254.1 illustrates carrier screening. Notice the tremendous reduction of the range of the Coulomb potential from carrier screening. The screened potential has the appearance of a square well.

The Debye screening length can be derived from solving the nondegenerate Poisson equation. We suppose that the potential contains two components: the potential of a constant background of charges and the potential from the spatially distributed electrons and holes whose concentrations vary spatially due to the impurity charge. Let the constant background charge be zero (charge neutrality),

$$\rho_o = q(P_o - N_o + N_{DD} - N_D) = 0$$

and let the potential be decomposed into two components, $V = V_0 + \delta V$. Let $\delta V << kT$ in order to linearize the problem. The reference $V_0=0$ will be used. The Poisson equation is then linearized.

$$-\nabla^2 V = \rho/\varepsilon_s \tag{254.2}$$

$$= (q/\varepsilon_s)(P - N + N_{DD} - N_D) \tag{254.2A}$$

$$= (q/\varepsilon_s)(P_o + \delta P - N_o - \delta N + N_{DD} - N_D) \tag{254.2B}$$

$$\approx (q/\varepsilon_s)(\delta P - \delta N) \tag{254.2C}$$

$$= (q/\varepsilon_s)(qV/kT)(P_o+N_o) \quad \text{(Boltzmann Approximation)} \tag{254.2D}$$

$$= V/r_s{}^2 \tag{254.2E}$$

where

$$r_s = \sqrt{\varepsilon_s kT/[q^2(P_o+N_o)]} \quad \text{(Debye screening length).} \tag{254.2F}$$

If the carrier density is high, Fermi statistics must be used and the Debye length is modified. In the strong degeneracy limit of very high carrier densities,

$$k_s{}^2 = (3/2)e^2N_o/(E_F-E_C) = 4(3N_o/\pi)^{1/3}/a_1 = e^2D(E_F) \tag{254.3A}$$

and

$$r_s = k_s{}^{-1} = ½\sqrt{a_1}(\pi/3N_o)^{1/6} \quad \text{(Fermi screening length).} \tag{254.3B}$$

The critical electron concentration at which the screened Coulomb potential has no bound state was found to be $k_s > 1.19/a_1$ or $r_s < a_1/1.19$. [See F.J.Roberts, H.C.Graboske,Jr. and D.J.Harwood, Physical Review A1, 1577 (1970).] Thus, there is an abrupt disappearance of the activation energy. The critical electron separation and electron concentration using the degenerate electron gas approximation are then

$$s_{cr} = 2.78a_1 = 38.4\text{A} \tag{254.4A}$$

and

$$N_{cr} = 0.0464/a_1{}^3 = 1.77\times10^{19}\text{cm}^{-3}. \tag{254.4B}$$

For nondegenerate gas, the results are

$$s_{cs} = [(qa_1/1.19)^2/9\varepsilon_s kT)]^{1/3} = 43.0\text{A} \tag{254.5A}$$

and

$$N_{cr} = \varepsilon_s kT(1.19/qa_1)^2 = 1.26\times10^{19}\text{cm}^{-3}. \tag{254.5B}$$

The exact solution at $r_s=0.82a_1=0.84\times13.8\text{A}=11.6\text{A}$ lies between these two approximations and is about $1.4\times10^{19}\text{cm}^{-3}$.

The above solution assumed that potential variation is small compared with kT. Although the potential energy curves of the bare and carrier-screened point charge in Fig.254.1 are reduced by the dielectric constant of the semiconductor ($\varepsilon_s/\varepsilon_0=11.8$ for Si), the resultant potential energy still varies much more than $kT \approx 25\text{meV}$. Thus, the Poisson equation cannot be linearized. The nonlinear problem must be solved to give a more accurate solution. In any case, it predicts an abrupt disappearance of both the bound state and thermal activation energy due to electron screening when the electron density is above a critical value. Thus, it can be tested experimentally, such as at low temperatures by injecting a high concentration of electrons or holes while using infrared absorption to detect the trapped electrons.

Because of the random spatial distribution of the impurity atoms in an otherwise crystalline semiconductor, there are enough electrons to screen the impurity atoms situated in regions of low impurity concentration. Thus, both (i) the impurity band formation or bound energy level shift due to the presence of surrounding impurities discussed in section 253, and (ii) carrier screening discussed in this section must be taken into account simultaneously. A nonlinear screened impurity cluster model must be solved to give a more accurate numerical condition.

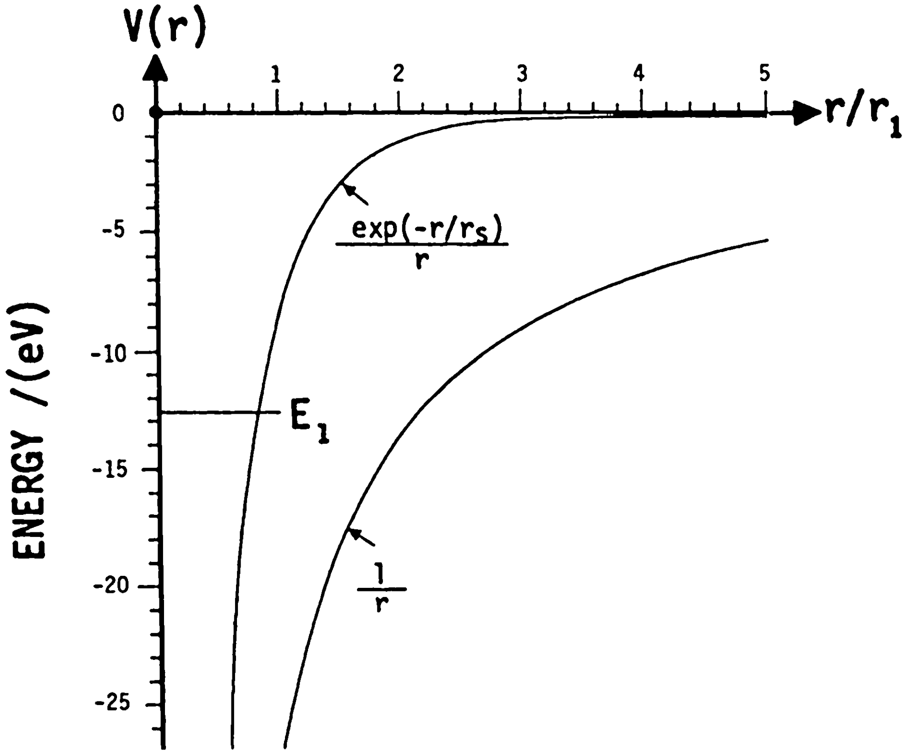


Fig.254.1 Effect of carrier screening on the electron potential energy in the Coulomb electrostatic force field of a positive point charge. The screening length is equal to the critical length, $r_s = a_1/1.19$, at which the Bohr bound states disappears.

299. BIBLIOGRAPHY

The following is a selected list of historical and current textbooks and references on the equilibrium properties of homogeneous semiconductor which complement and some of which extend the mathematical details and physics coverage just given in chapter 2. However, the fundamental considerations of the state of equilibrium and various partial equilibria described in chapter 2 have not been covered as a group in the literature. The sequence follows roughly that used by this author when he began to gather materials for teaching an undergraduate and a graduate course on semiconductor devices and physics in 1961. Some of the books in last part of this list extend the coverage to intermediate and advanced levels for a senior and first graduate course on equilibrium and nonequilibrium statistical mechanics.

[299.1] William Shockley (Bell Telephone Laboratories, Shockley Transistor Corporation of Beckman Instruments, Inc., Stanford), **Electrons and Holes in Semiconductors**, D. Van Nostrand Company, Inc. New York, 1950. In one short chapter 16 (pp.452-480), Shockley gave the essential physics, the derivation by counting the number of ways of arranging a group of particles, and the relevant semiconductor-device application examples of the energy distribution of three type of quasi-particles at thermodynamic equilibrium: the electrons (and holes), the phonons, and the trapped or bound electrons (and holes). Phonons are quantized lattice vibration waves to be described in chapter 3. The electron quasi-particle (spin=1/2) and other quasi-particles and 'real' particles with an odd-integer spin number are also known as **Fermions** because their number distribution in energy or kinetic energy is governed by the **Fermi** or **Fermi-Dirac** distribution law. The phonon quasi-particle (quantized lattice vibrations with spin=0) and other quasi-particles and 'real' particles (photons and others) with an even-integer spin number are also known as **Bosons** because their number distribution in energy or kinetic energy is governed by the **Bose** or **Bose-Einstein** distribution law. The trapped electrons or bound Fermions is governed by the **trap** distribution law which was first included by Shockley in a book.

[299.2] Francis Weston Sears (M.I.T.), **An Introduction to Thermodynamics, the Kinetic Theory of Gases, and Statistical Mechanics**, Addison-Wesley Press, Inc. Cambridge, MA, 1950. This is still the best introductory textbook on the Fermi, Bose, Boltzmann and Maxwellian distributions and their fundamental connections with the temperature and other thermodynamic parameters (the entropy and the various thermodynamic energies). Counting the number of ways of arranging particles is the model used for the derivation. Having already learned thermodynamics in freshman chemistry and sophomore physics, the students can go directly to chapter 11 to begin a study on the equilibrium classical and quantum statistical and kinetic theory of electrons.

[299.3] John P. McKelvey (Pennsylvania State University), **Solid State and Semiconductor Physics**, Harper & Row, Publishers, New York, 1966, see also the 1990 edition. This is one of the first junior-senior textbook specially for developing the fundamental physics concepts needed for device engineers and device physicists, including statistical mechanics of electrons, phonons, photons, and trapped electrons. Derivation and examples are given in chapter 5 on the Fermi-Dirac, Boltzmann-Maxwellian, and Bose-Einstein distributions, and in chapter 9 on the trap distribution.

[299.4] J. S. Blakemore (Honeywell, Florida Atlantic University, Oregon Graduate Research Center), **Semiconductor Statistics**, Pergamon Press, 1962. This is the first and still the only book that gives many application examples of the Fermi-Dirac as well as trap distribution functions, including many-electron traps. It does not give the derivation of the distribution functions, only applications.

[299.5] F. Reif (Berkeley), **Fundamentals of Statistical and Thermal Physics**, McGraw-Hill Book Co. 1965. Chapter 9 gives in-depth advanced-level derivations (using partition functions) of the Fermi and Bose distributions, and some applications including fluctuation from equilibrium. As a typical physics book, it ignors the trap statistics of trapped or bound electrons, and the statistics of particles each trapped by a potential well.

299. PROBLEMS

P202.1 Describe the partial equilibrium conditions of a piece of Si sitting on top a desk in a room at a space-time constant temperature of 300K.

P202.2 A piece of pure silicon is immersed in pure boiling water. What are the essentially equilibrium parameters and the dominant nonequilibrium parameters?

P202.3 A solar cell panel is attached to the exterior of an orbiting satellite which can be considered a vacuum. The panel is initially in the shadows. It is suddenly exposed to the sunlight. Describe the transient events that lead to thermal equilibrium of the atoms and the remaining nonequilibrium events. Disregard the electrons and holes for this problem. Consider also the electronic equilibria after optoelectronic devices are studied.

P202.4 Extend the considerations of microscopic and macroscopic nonequilibrium and equilibrium of isolated and interacting-contacting material bodies given in this section to an everyday environment or situation, and to the universe.

P202.5 What is the de Broglie wavelength of an electron at the average kinetic energy k_BT at $T=300K$? [Why not $(3/2)k_BT$?] This is known as the **thermal de Broglie wavelength** of an electron. How many valence electrons are there in a thermal de Broglie cube of Si whose edge is the thermal de Broglie wavelength? What is the fluctuation in electron number in such a cube? What is the significance of localization of an electron by a wave packet in a thermal de Broglie cube?

P202.6 A piece of crystalline silicon slice contains an impurity whose concentration is given by $N_{Impurity}=ax$ where x is alone the direction perpendicular to the surfaces of the Si slice and a is a constant. The Si is held motionless and immersed in a stationary ambient gas. The two surfaces of the Si slice are in perfect thermal contact with the ambient gas which is held at 300K and which can be considered as an infinite heat sink and heat source. Is the Si slice at thermodynamic equilibrium? Does it approach several partial equilibrium conditions? Give reasons.

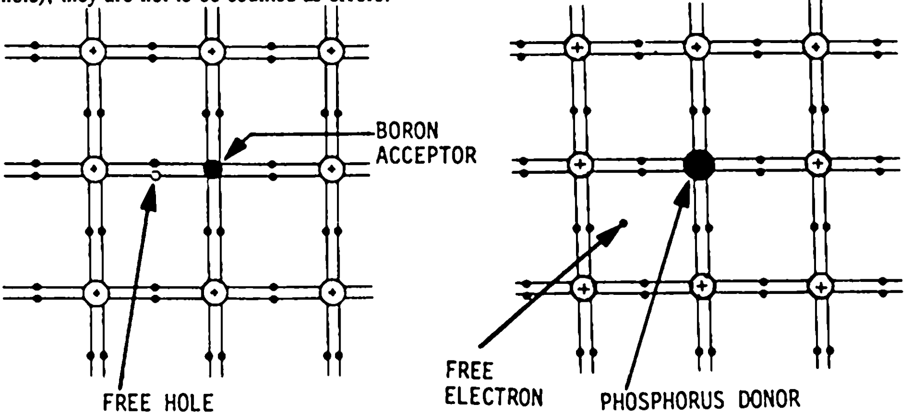
P210.1 How many intrinsic electrons and holes are there in a thermal de Broglie cube at 300K in Si corresponding to an energy equal to k_BT? What is the thermal fluctuation in carrier density in this cube? Can the intrinsic electrons be localized?

P223.1 The binding energy of an electron trapped to the substitutional As donor in Si was measured spectroscopically (infra-red absorption at $T=4.2K$) in 1965 by Prof. Ramdas and his graduate students at Purdue and found to be 53.69meV. Assume that the static dielectric constant is valid for screening the As ion charge by the valence electrons and is given by $\varepsilon_s/\varepsilon_0=11.8$. What is the orbital radius of an electron trapped at the ground state by As? How many valence electrons are enclosed in the Bohr sphere or in the spherical volume enclosed by the trapped electron orbit? Is the assumption on the static dielectric constant reasonable and why?

P223.2 The binding energy of an electron trapped to a sulfur double donor S^{+6} in Si was measured by Prof. Sah and his graduate students at Illinois in 1971 using the photocapacitance transient method they invented and found to be 613.26 meV. Assume that the static dielectric constant is valid for screening the As core charge by the valence electrons and is given by $\varepsilon_s/\varepsilon_0 = 11.8$. What is the orbit radius of an electron trapped at the ground state? How many valence electrons are enclosed in the Bohr sphere or in the spherical volume enclosed by the trapped electron orbit? Is the assumption on the static dielectric constant reasonable and why?

P223.3 The thermal activation and photo-excitation energies of one and two electrons trapped to the double or 2-electron sulfur donor were measured by Sah and his graduate students in 1971 by observing the photo and dark capacitance transients in a sulfur doped silicon p/n junction diode. They found that the activation energies are 613.26 meV and 302.0 meV. Draw the energy band and transition energy diagrams for one and two electrons trapped to a double sulfur donor center.

P223.4 Give the basic physics reason of at least three errors in each of the two bond diagrams of shallow acceptor and donor in semiconductors shown below. These errors are frequently found in freshman chemistry, sophomore physics, and junior semiconductor device textbooks as well as reference books. Mixed symbols are used: Shockley (two bars), Lewis (dots), and ours (circle for hole); they are not to be counted as errors.



P231.1 Why is it that we can write the Boltzmann factor in terms of the total energy, E, such as exp(-βE), while the basis given in the text discussion of this chapter suggests that it should be written in terms of the kinetic energy (KE), $m^*v^2/2$ or $\hbar^2k^2/2m^*$, such as exp(-$\beta m^*v^2/2$) or exp(-$\beta h^2k^2/2m^*$)?

P232.2 Why is the kinetic energy of electrons measured from E_C and holes from E_V rather than vacuum level? Hint: Apply similar reasoning for the binding energy E_I near the end of section 223.

P231.3 Connect the Boltzmann approximation of the Fermi function to the Maxwellian velocity distribution function given in your freshman chemistry textbook. If your freshman chemistry book does not have an introductory discussion of statistical mechanics and kinetic theory of gases, find a textbook for the honors section or for chemistry major which does.

P231.4 Give the reasons why the following commonly and carelessly made statement is erroneous. "The principle of detailed balance states that in thermodynamic equilibrium every process and its inverse process proceed at equal rates." While this statement is fundamentally deficient, the equality of (231.1) and (231.2) asserted and explained in section 231 is correctly stated and fundamentally flawless.

P231.5 Which of the following statements are fundamentally correct and which ones are fundamentally incorrect if taken at their face value without implied interpretation? Give the physics reasons. (a) Half of the energy levels at E_F is occupied. (b) The energy level at E_F is occupied by one electron. (c) The probability that the energy level at E_F is occupied is 50%.

P231.6 Obtain the correction terms to fourth order in energy (or E^4) in the Boltzmann approximation by expanding the Fermi function in the Taylor series. $\exp(Z) = Z^0/0! + Z/1! + Z^2/2! + Z^3/3! + Z^4/4! + \ldots + Z^n/n! + \ldots$ where n=integer or 0. How do the absolute and percentage errors vary with energy?

P231.7 Prove in one line of algebra that the hole occupation function $f_p(E_p)$ is identical in form to the electron occupation function, $f_n(E_n)$, and they both have the same form of the Fermi function, f(E), if the independent variable, E, is the particle energy, i.e. both E and E_F have the same sign as the particle potential energy. Namely, prove that for holes, $f_p(E_p) = 1/\{1+\exp[(E_p-E_{Fp})/kT]\}$ which has exactly the same form as that for electrons, $f_n(E_n) = 1/\{1+\exp[(E_n-E_{Fn})/kT]\}$ or as the Fermi function for electrons, $f(E) = 1/\{1+\exp[(E-E_F)/kT]\}$ derived section 231.

P231.8 In what energy range (or above what energy) will the Boltzmann approximation give less than 1% error (a) in electron occupation, and (b) in hole occupation? Use the Fermi energy, E_F, as the reference.

P231.9 At what energy is the Fermi function equal to 1/2? Calculate and tabulate the occupation probability of the energy levels at $(E-E_F)/kT = -5, -4, -3, -2, -1, 0, +1, +2, +3, +4, +5$ (a) by electrons, f(E), and (b) by holes, 1-f(E). (c) Tabulate also the probability and (d) the percent error of both electron and hole occupation factors if the Boltzmann approximation is made. Does the solution obtained in Problem P231.5 agree with the table?

P231.10 What is the fraction of the electron states occupied by electrons at an energy of $E=E_F+0.0455$eV at T=300K? Use the very accurate value of the Boltzmann constant. What is the fraction of these states occupied by holes (or empty of electrons)? What are the electron and hole occupancy fractions if the Boltzmann approximation is made and what are the percentage errors? Explain why one error is negligible while the other is significant?

P232.1 A pure and very thin Si slice is placed inside of and in parallel with a parallel-plate capacitor which has a spatially constant electric field. The electron potential energy inside the Si varies linearly with distance in the thickness direction, V(x)=ax where $a=2\text{x}10^7$V/cm$=5^{-1}$(V/A) where A=Angstrom unit$=10^{-8}$cm. Sketch to scale the E-x energy band diagram of this Si. Plot (below the energy band diagram) the electron and hole concentration N(x) and P(x) from x=-25A to +25A at 21.2C ($n_i=10^{10}\text{cm}^{-3}$). Plot the concentrations first in semilog scale (you get straight lines) then read off a few points from the semilog plot to construct the linear plot. What impression do you get on the rapidity of the variation of the electron and hole concentrations with position? State quantitatively.

P233.1 How does the volume density of electron state per unit range dk, D(k), and per unit energy range dE, D(E), vary with k and E in (a) 1-d, (b) 2-d and (c) 3-d crystal? Assume the simple parabolic, hyperbolic and spherical energy bands respectively: (Partial incomplete answers are given.)

1-d $E-E_C=(\hbar^2/2m)(k_x^2)$; $D_1(k_x)=(1/\pi)$ $D_1(E)=C_1/\sqrt{(E-E_C)}$
2-d $E-E_C=(\hbar^2/2m)(k_x^2+k_y^2)$; $D_2(k)=(k/\pi)$ $D_2(E)=C_2=m/(\pi\hbar^2)$
3-d $E-E_C=(\hbar^2/2m)(k_x^2+k_y^2+k_z^2)$. $D_3(k)=(k/\pi)^2$ $D_3(E)=C_3\cdot\sqrt{(E-E_C)}$

P233.2 A one-dimensional hypothetical crystal made of hydrogen atoms is 1 cm long, contains 10^8 hydrogen atoms each has one valence electron, and has an interatomic spacing of 1A or 10^{-8}cm (for numerical simplicity while in real crystals, it is 2-5A). (a) What is the number of energy levels in the 1s conduction band? (b) What is the density of electron energy levels in 1% of k-space, i.e. $dk=0.01\times(2\pi)$ and what is the density of electron states in this dk interval? (c) If the electrons can be approximated by free electrons, $E=\hbar^2k^2/2m$, how does the density of state per unit energy range vary with energy? (d) Show that C_1 in P233.1 is $\sqrt{2m}/\pi\hbar$. (e) What is the numerical value of C_1 in $(\text{cm-eV})^{-1}$ when $E=E_C+1\text{eV}$ if $m=m_{free}$? (f) What is the position of the Fermi level (in eV) at T=0K measured relative to the bottom of the conduction band? (g) What is the wave number at the Fermi energy $k_F(\text{cm}^{-1})$? (h) What is the density of states per unit wave number range at the Fermi energy, $D_1(k_F)$, in units of state? (h) What is the density of states per unit energy range at the Fermi energy, $D_1(E_F)$, in units of state/cm-eV?

P233.3 Repeat P233.2 for a 1-d helium crystal (2 valence electrons per host atom).

P233.4 Repeat P233.2 for a 2-d hydrogen crystal in a square lattice.

P233.5 Repeat P233.2 for a 2-d helium crystal in a square lattice which has two valence electrons per host atom.

P233.6 Repeat P233.2 for a 3-d hydrogen crystal in a cubic lattice.

P233.7 Repeat P233.2 for a 3-d helium crystal in a cubic lattice (2 valence electrons per host atom).

P233.8 For a crystal of finite size it is found that there are 100 allowed electron energy levels at the energy E_1 in the conduction band and that the Fermi energy lies at E_1, i.e. $E_F=E_1$. How many electrons are there whose energy is E_F?

P233.9 For a crystal of finite size, it is found that there are 100 allowed electron energy levels at an energy E_F. Which of the following statements are correct and which are wrong, and why? (a) Each of the 100 levels is occupied by one electron. (b) 50 levels are empty and 50 levels are each occupied by two electrons. (c) The first 50 levels are empty and the last 50 levels are each occupied by two electrons. (d) The probability that each level is occupied by one electron is 0.5. (e) 100 states are occupied by one electron each and 100 states are empty.

P233.10 When the Boltzmann approximation is valid for computing the electron concentration, is it also valid for computing the hole concentration? Give two examples, one showing the true and the other, the false answer.

P233.11 The electron concentration, $N=N_C\exp[-(E_F-E_C)/kT]$, and hole concentration, $P=N_V\exp[-(E_C-E_F)/kT]$, were derived based on four assumptions. What are these assumptions? Why are they made? And how are they justifiable? Partial answers: large crystal, spherical energy band, large band width, Boltzmann. For engineering inclined students who want to get the analytical expressions and accurate numerical criteria for these four approximations, do the following four problems or look up their solutions in an advanced semiconductor physics textbook. The derivations require only straight forward algebra and elementary mathematics and functions.]

P233.12 Using the Taylor series expansion for the Fermi function, $f(Z) = 1/(1+Z) = 1-Z/1! + Z^2/2! - Z^3/3! + ...$, where $Z=(E_C-E_F)/kT$, obtain the correction terms to fourth order of Fermi energy in the bracket $\{1+...\}$ of the electron concentration expression, $N = N_C\exp[-(E_C-E_F)/kT]\cdot\{1 + F_1Z + F_2Z^2 + F_3Z^3 + F_4Z^4 + ...\}$, i.e., obtain the analytical expression for F_1, F_2, F_3, F_4. Compute the Fermi energy position above which the electron concentration from the Boltzmann approximation is in error by 0.01%, 0.1%, 1%, and 10%.

P233.13 Give the mathematical criterion for the wide bandwidth approximation used to derive the electron or hole concentration expression given in Problem P233.11, i.e. obtain the analytical or series solution of the Fermi integral from $\varepsilon=\varepsilon_{c'}-\varepsilon_c$ to ∞. Find the minimum bandwidth, $BW=\varepsilon_{c'}-\varepsilon_c$, for a maximum error of 0.01%, 0.1%, 1% and 10% in the Boltzmann approximation of the electron concentration.

P233.14 What is the correction term or terms to the Boltzmann approximation of the electron or hole concentration if the next higher order term of the energy surface in k-space, i.e. the nonparabolic or cubic term, is taken into account. $E=(\hbar^2k^2/2m^*)\cdot\{1+k/k_3+...\}$. Find the expression to give $\Delta\%$ error in the Boltzmann approximation of N in terms of the nonparabolicity coefficient k_3. Give the numerical value of k_3 at 1% error if $k_3=1000(2\pi/a)$. Plug in the lattice constant for Si.

P233.15 What is the error in the Boltzmann approximation if the crystal size is not infinite but contains 10^2, 10^3, 10^4, 10^5, 10^6, 10^9 atoms? This is a very real situation in submicron transistors in which an active region has a very small volume and one must choose a smaller differential volume dxdydz in order to give many volume elements in a finite difference analysis of the problem. For example, in an Si volume of $0.1\text{x}1\text{x}1\mu m^3=10^7A^3= 10^{-17}cm^3$, such as the base of a BJT or channel of a MOST, there are only $10^{-17}\text{x}2\text{x}10^{22}=2\text{x}10^5$ Si atoms. This base or channel layer must be partitioned into 10^2 or more volume elements in order to accurately calculate the transistor characteristics using numerical techniques. Then, each volume would have only 2000 Si atoms.

P233.16 Find the position of the Fermi energy level at T=0K if the electron concentration in an n-type Si is 10^{15}, 10^{16}, 10^{17}, 10^{18}, 10^{19}, 10^{20}, or 10^{21} cm^{-3}. At what electron concentration, is the Boltzmann approximation no longer valid and give the reason. (A more detailed quantitative treatment is given later, in section 251. An estimate suffices in this problem using the approximations described in section 233.)

P233.17 Repeat P233.16 for the 1-d and 2-d hydrogen crystal. You need to decide the decade ranges of linear and areal electron concentration you must use to demonstrate the key ideas that are asked in P233.16.

P233.18 Suppose that there is an energy gap separating the 1s and 2s bands of the Helium crystal which is E_G. (See Fig.182.5.) Where is the Fermi level at T=0K if the electron concentration is n_v+N where N is the added electron density while n_v is the valence electron density. Compute the E_F position for the 1-d and 2-d Helium crystal for the appropriate decade ranges of electron concentrations similar to those in P233.17. (Hint: Use the band filling rules described in section 173.)

P241.1 The experimental formulae for the intrinsic carrier concentration in silicon was obtained by Morin and Maita at the Bell Telephone Laboratories in 1954 from conductivity and Hall voltage measurements over a wide range of temperature in many n-type and p-type impurity doped silicon single crystals. Their results were fitted to give $n_i^2=1.5\text{x}10^{33}T^3\exp(-1.21eV/kT)$ where k is the

Boltzmann constant given by (231.3) ($k=8.616\times10^{-5}$ eV/K, verify k in this unit) and T is the absolute temperature in Kelvin. Calculate the value of n_i at T=300°C and verify your result with that given by Fig.241.1(a).

P241.2 Where is the Fermi level in a pure Si at T=0K, 300K, and 600K. Use the density of state effective masses given in section 233 following (233.10B).

P242.1 The mass action law introduced in section 242 to compute the electron and hole concentrations and the principle of detailed balance introduced in section 231 to derive Fermi-Dirac distribution function of the electrons are directly related, one leads to the other, and yet are different at the fundamental level. This difference is frequently overlooked by semiconductor physicists, transistor engineers, and teachers. Describe the fundamental differences and show which one is more basic and leads to the other. (The derivation of the N and P using f(E) makes the answer obvious and trivial. But you need to describe your model based on an ensemble of electrons in dxdydxdE at E and (x,y,z) to make the description fundamentally rigorous.)

P242.2 Compute the thermal equilibrium constant K of the electron-hole pair generation-recombination reaction in silicon at 300K, 600K and 900K.

P242.3 A test charge is generated inside a pure Si at T=300K. How long does it take for the test charge to decay to 10% of its original magnitude? Where does the test charge go and give the reasons? Describe carefully what is meant by a test charge - its volume, the number of electrons it contains, etc.

P242.4 In the middle of section 242 it was demonstrated that the components of a carrier concentration cannot be added to give the total carrier concentration. Give the reason and demonstrate mathematically why the following result is incorrect even though electrical neutrality is satisfied. Assume complete ionization in an n-type extrinsic semiconductor doped with only one donor species of N_{DD} concentration. $N = N_{DD}$(from donors) + n_i(from breaking covalent bonds) and $P = n_i$(from breaking covalent bonds). Thus, $P-N-N_{DD}=0$ or electrical neutrality is satisfied.

P242.5 What is the equilibrium concentration of electrons and holes in an n-type silicon at T=21.2°C ($n_i=1.0\times10^{10}\text{cm}^{-3}$) in which $N_{DD}=4\times10^{10}\text{cm}^{-3}$ and are all ionized, and $N_{AA}=0$? How much error is introduced if the intrinsic electron is neglected?

P242.6 A homogeneously doped n-type Si at 21.2°C ($n_i=1.00\times10^{10}\text{cm}^{-3}$) has $N_{DD}=1.25\times10^{15}\text{cm}^{-3}$ donors and $N_{AA}=0.25\times10^{15}\text{cm}^{-3}$ acceptors, all of which are ionized. At some instance of time, for example t=13:00, the carrier concentrations were found to be $N=1.25\times10^{15}\text{cm}^{-3}$ and $P=0.25\times10^{15}\text{cm}^{-3}$ and were uniform over the entire volume of the crystal. (a) Is the crystal electrically neutral at t=13:00? (b) Why is the crystal not at thermodynamic equilibrium at t=13:00 and back up your answer quantitatively?

P243.1 A homogeneously impurity doped silicon has $N_{DD}=1.01\times10^{16}$ phosphor/cm^3 and $N_{AA}=1.00\times10^{14}$ boron/cm^3, all of which are ionized. What are the electron and hole concentrations at 21.2°C and 362°C. Use Fig.241.1(a) or the experimental formulae given in problem 241.1 to obtain the accurate value of n_i at these two temperatures.

P244.1 What is the intrinsic temperature of an n-type silicon doped with $N_{DD}=10^{18}\text{cm}^{-3}$?

P244.2 Above what dopant impurity concentration will a silicon become extrinsic at a temperature just below its melting point? The theoretically extrapolated intrinsic carrier concentration at the melting point is about $5\times10^{20}\text{cm}^{-3}$. Is this physically possible for germanium?

Use a reasonable extrapolation to get the intrinsic carrier concentration of Ge at its melting point. The energy gap of a semiconductor decreases with increasing temperature due to the high intrinsic concentration of electrons and holes which screens and decreases the electron-hole Coulomb attraction force and hence lowers the covalent bond energy to break a bond or the energy gap. However, the energy gap increases due to the larger interatomic distance or lattice constant caused by thermal expansion at higher temperatures. But this increase in E_G is much smaller than the decrease due to screening. A theoretical extrapolation of the experimental data by Sah in 1966, and described in an advanced semiconductor physics textbook, gave E_G just below melting point of $E_G(\mathrm{Si},1412\mathrm{C})=0.73\mathrm{eV}$ and $E_G(\mathrm{Ge},T=957^\circ\mathrm{C})=0.30\mathrm{eV}$.

P244.3 A semiconductor contains a p-type region doped with boron acceptor and an n-type region doped with phosphorous. The boundary between the p-type and n-type regions is known as the p/n junction whose electrical properties will be studied in detail in chapter 5. This is an electrical boundary since there is no physical or metallurgical boundary between the p-type and n-type regions because only the very-low-concentration (compared with 2×10^{22} Si/cm^3) dopant impurity type is changed from boron to phosphorus. What is the maximum temperature when the electrical boundary disappears if $N_{AA}=10^{18}\mathrm{cm}^{-3}$ and $N_{DD}=10^{15}\mathrm{cm}^{-3}$?

P244.4 Use the transition energy band diagram of a semiconductor containing N_{DD} cm^{-3} donors and N_{AA} cm^{-3} acceptors to give the physical reasons why the two components of the electron concentration, intrinsic from thermal breaking of the covalent bonds and extrinsic from thermal release of impurity electrons, cannot be added algebraically in the extrinsic temperature range as explained in (242.10) but which can be added algebraically near and above the intrinsic temperature as explained in section 244 and given by (244.1C). This is an extension of using the energy band diagrams Fig.244.1(a) and (b) for temperatures above T_I.

P251.1 Expand the Fermi-Dirac integral (251.2) in power series of η in two ways: direct and expansion of $\exp(\varepsilon-\eta)$ like P233.12. What are the differences between the two? Use the general definition of the Fermi-Dirac integral of j-th order

$$F_j(\eta) = \frac{1}{\Gamma(1+j)}\int_0^\infty \frac{\varepsilon^j d\varepsilon}{1+\exp[\varepsilon-\eta]} \qquad (251.2)$$

and their derivatives are given by

$$dF_j(\eta)/d\eta = F_{j-1}(\eta)$$

These functions are tabulated. [For the tabulations, see references given by Blackmore, Solid-State Electronics, 25(11), 1067, (1982).]

P251.2 Determine the first order error term in the nondegenerate mass action law $NP=n_i^2$ which was derived at low carrier concentrations. Assume (a) only one carrier density (electron or hole but not both) is nearly degenerate and (b) both carrier densities are nearly degenerate. Use the Blackmore approximation for the Fermi-Dirac integral, (251.2B). Compare the analytical result in (a) with the more rigorous expansions from Problems P233.12 and P251.1. Put in numbers for Si at $T=300\mathrm{K}$ and $N=10^{17}$, 10^{18}, 10^{19} and $10^{20}\mathrm{cm}^{-3}$.

P252.1 A representative volume element dxdydz of a uniformly doped n-type silicon crystal contains 100 substitutional phosphorus donors distributed on the superlattice so there is no local fluctuation of positions. Which of the following statements are correct and incorrect? Give the fundamental reasons. (a) f_D is the fraction of the 100 donors each with one trapped electron. (b) f_D is the fraction of neutral donors. (c) $1-f_D$ is the fraction of the 100 donors unoccupied by electron. (d) $1-f_D$ is the fraction of donors each with one trapped hole. (e) $1-f_D$ is the fraction of ionized donor. (f) When $E_F=E_D$, half of the donors are neutral. (g) When $E_F=E_D-kT\log_e g_D$, half of the donors are ionized. (h) When $f_D=1/2$, the first 50 donors are each occupied by an electron. (i) When $f_D=1/2$, 50 of the 100 donors are each occupied by a spin-up electron. (j) When $f_D=1/2$, 50 of the 100 donors are each occupied by an electron, some have spin up and others have spin down. (k) When $f_D=1/2$, about 25 donors are each occupied by a spin-up electron and the other 25, a spin-down electron. (l) When $f_D=1/4$, 25 donors are each occupied by an electron and 75 donors are each occupied by a hole.

P252.2 It was proposed and demonstrated, using the gold levels in Si, by this author in the mid-1970 in a research project on solar cell efficiency for the Department of Energy and Jet Propulsion Laboratory that by carefully analyzing the experimental data to give an experimental value of the degeneracy factor, g, for the trapped electron or trapped hole, one can use the difference in f_D and f_A to determine whether the observed trap is an electron or hole trap. Demonstrate this possibility by showing that $f_D(E_D-E_F,g_D)\neq 1-f_A(E_F-E_A,g_A)$. Suppose that by Hall effect and dark capacitance transient measurements, a negatively charged trap (acceptor-like by definition of section 222) in an n-Si is found to have $g_T(\text{data})=12$ and $E_C-E_T=0.15\text{eV}$. Is it an electron trap or a hole trap and why? (Answer: electron trap.)

P252.3 Give the fundamental reasons on the symmetry between the trapped electron and trapped hole distribution functions given by (252.3) and (252.4) respectively.

P252.4 In the examples given in section 252 on impurity deionization, it was stated in the text that the problem can be solved analytically using either the brute force method or the mass action law. The algebra via the brute force route was presented. Derive the analytical solution using the mass action law applied to both of the two reactions: (i) the thermal generation-recombination of electron-hole pairs across the energy gap, $NP = K_i = n_i^2$, and (ii) the electron trapping-detrapping or capture-emission at the donors, $N(N_{DD}-N_D)=K_DN_D$. Show that only at low temperatures or in a not-too-narrow-gap semiconductor and/or at low donor concentrations is Henry's law true, i.e., K_D is independent of carrier concentration.

P252.5 Using only basic physics, no algebra, and no equations, show that at $T=0K$, $E_F=(E_C-E_D)/2$ in an n-type semiconductor with $N_{DD}=10$ and $N_{AA}=0$.

P252.6 Using only basic physics, no algebra, and no equations, show that at $T=0K$, $E_F=(E_C-E_D)$ in an n-type semiconductor with $N_{DD}=10$ and $N_{AA}=1$.

P252.7 This and the next problem summarize the equation and methodology used by physicists to calculate N_{DD}, N_{AA}, E_D and g_D in a semiconductor from a set of N versus T data from very low temperatures to above T_i. Using the mass action law for all the equilibrium reactions present in this problem, derive the general cubic equation for the electron concentration in a nondegenerate semiconductor that has $N_{DD} > N_{AA} \neq 0$ which is valid at all temperatures. Obtain the quadratic equation and the explicit analytical solution for N which is valid at all temperatures below the intrinsic temperature. Answers: $(N+K_D)[N+N_{AA}-(n_i^2/N] = K_DN_{DD}$, $(N+K_D)(N+N_{AA}) \propto K_DN_{DD}$, and $N = \tfrac{1}{2}(N_{AA}+K_D)\{[1 + 4K_D(N_{DD}-N_{AA})/(N_{AA}+K_D)^2]^{1/2} - 1\}$.

P252.8 Show analytically that in the intermediate low temperature range, $T_2< T < T_4$, the electron concentration can be approximated by

$$N = \sqrt{N_C(N_{DD}-N_{AA})}\exp[-(E_C-E_D)/2kT]$$

and at the lower temperatures, $T < T_4$, the approximate electron concentration is

$$N = [N_C(N_{DD}-N_{AA})/2N_{AA}]\exp[-(E_C-E_D)/kT].$$

Derive the expression for T_1, T_2, T_3 and T_4. In $T_i<T<T_1$, $N=N_{DD}-N_{AA}$, and at T_3, $K_D(T_3)=N_{AA}$ and $N_3=\surd N_{AA}(\surd N_{DD} - \surd N_{AA})$. Find also the analytical expressions for N_1, N_2 and N_4. Mark these on the experimental data of the p-type Si given in Fig.252.3.

P253.1 Phosphorous donor bound state disappears when the phosphorus concentration reaches the critical value N_{sc} according to (253.5) which gives a numerical value of $7.43\times10^{17}\text{cm}^{-3}$. How does the electron concentration vary with temperature if $N_{DD}=10^{18}\text{cm}^{-3}$. Sketch this in Fig.252.1(a) to compare with the persistent bound-state-theoretical curves given in Fig.252.1(a). Give the reason (in terms of location fluctuation of the donor atoms) to account for the residual temperature dependences when $N_{DD}>N_{cr}$ shown in the data of Fig.253.1. Modify your curve to take this local fluctuation of donor impurity concentration into account by assuming that there are two groups of donors. One group with concentration $N_1>N_{cr}$ screens out the donor bound state and hence will conduct. The other group (in the impurity spatial distribution tail) with $N_2<N_c$r will be trapped at E_D and hence will not conduct. Plot N_1 and N_2 from some experimental curves given in Fig.253.1.

P254.1 Assume that carrier screening from impurity electrons will make the impurity bound state disappear by disregarding the one-electron per one-impurity difficulty (the idea of screening by the electron density fluctuation rather than by individual electron). (a) How will the theoretical N versus 1/T be modified at the higher temperature ranges in a sample that has 10^{20} phosphorus/cm^3? Next, assume that there are two groups of electrons with concentrations $N_1<N_{cr}$ and $N_2<N_1$, then the N_2 bound states will be screened out by the N_1 electrons when their concentration N_1 reaches N_{cr}. (b) How does the presence of the second group of electrons affect the N versus 1/T curve?

P255.1 Sketch the potential energy or E_C and E_V as a function of position in a semiconductor crystal which contains (a) a complete random distribution of a small fraction of the host atoms from their lattice sites, (b) non-crystalline or amorphous regions where the host atoms are completely randomly located, (c) donor impurity atoms located on a superlattice, (d) substitutional donor impurity atoms completely randomly distributed, (e) substitutional donor impurity atoms not as randomly distributed as in problem (d), (f) donor-acceptor impurity pairs whose separation is completely randomly distributed.

Chapter 3

DRIFT, DIFFUSION, GENERATION, RECOMBINATION, TRAPPING and TUNNELING

300 INTRODUCTION 232
310 DRIFT 233
  311 Drift Velocity in an Electric Field, 233
  312 Drift Current, Drift Mobility and Conductivity, 238
  313 Temperature Dependences of the Drift Mobility, 239
    * Ionized Impurity Scattering, 240
    * Phonons for Describing Lattice Scattering, 241
    * Lattice Scattering, 246
  314 Electric Field Dependence of Mobility, 251
  315 Intrinsic and Extrinsic Conductivities of a Semiconductor, 252
    * Intrinsic Conductivity of Pure Semiconductor, 253
    * Conductivity of Impure Semiconductor, 253
320 DIFFUSION 254
  321 The Einstein Relationship, 259
  322 The Boltzmann Relationship, 259
  323 Examples of Diffusion Current, 260
330 CONSTANCY OF THE FERMI ENERGY LEVEL 261
  331 The Quasi-Fermi Levels and Quasi-Fermi Potentials, 262
340 CONTINUITY EQUATION OF CHARGE AND CURRENT 265
350 THE SHOCKLEY EQUATIONS OF SEMICONDUCTORS 268
360 GENERATION, RECOMBINATION, TRAPPING, AND TUNNELING 270
  3611 Interband Thermal Generation and Recombination, 273
  3612 Interband Optical Generation and Recombination, 275
  3613 Interband Auger Recombination and Impact Generation, 278
  3621 Band-Trap Thermal (SRH) Generation-Recombination-Trapping, 281
  3622 Band-Trap Optical Generation-Recombination-Trapping, 285
  3623 Band-Trap Auger Capture and Impact Emission, 286
  363n Three Intertrap Transitions, 286
  36n0 Elastic Tunneling, 287
  36n4 Inelastic Tunneling, 289
  36n5 Collective Transitions, 290
370 LIFETIMES 291
  371 Interband Thermal and Optical Recombination Lifetimes, 291
  372 Band-Trap Thermal (SRH) and Optical Recombination Lifetimes, 294
  373 Lifetimes for Simultaneous Presence of Many GRTT Mechanisms, 295
380 PHYSICS AND DATA OF THE GRTT RATE COEFFICIENTS 296
  381 Thermal (SRH) Capture and Emission Rates, 296
  382 Optical Emission Rate, 300
  383 Interband Optical Generation Rate, 300
  384 Interband Impact Generation Rate, 302
  385 Interband Tunneling Rate, 303
  386 Band-Trap Tunneling, 305
399 BIBLIOGRAPHY AND PROBLEMS 306

300 INTRODUCTION

Chapter 1 introduced the electron-pair bond and electron energy band models of solids. The nuclei of the host atoms were assumed at rest (absolute zero temperature) and electrons seek the lowest energy levels allowed by Pauli's exclusion principle. Chapter 2 described the effects of atomic or nuclear vibration, characterized by a finite temperature, on the kinetic energy distribution of the electrons at thermal (or thermodynamic) equilibrium between the electrons, holes and atomic cores. Thermal equilibrium is characterized by space-time constancy of the average kinetic energy of each particle species (electrons, holes, vibrating nuclei or phonons). The constancy is maintained by the frequent interparticle collisions via the Coulomb force. Thermal equilibrium is characterized mathematically by the Fermi (or Fermi-Dirac) distribution function of the kinetic energy of the electrons and the Bose (or Bose-Einstein) distribution of the kinetic energy of the vibrating nuclei or phonons.

In this chapter, chapter 3, the physical mechanisms causing nonequilibrium and net transport of electrons and holes are described. The mathematical formulae of nonequilibrium are derived in order to model, characterize, and predict the electrical currents in diodes and transistors in later chapters. The fundamental cause of nonequilibrium is the exposure of the solid to an **applied force**, via interparticle collisions at the solid/solid, solid/gas, solid/liquid interfaces and by action at a distance described by electric, magnetic or electromagnetic fields.

The fundamental nonequilibrium or transport mechanisms can be put into **five** fundamental categories in terms of the motion of electrons: **drift** under an electric field, **diffusion** due to an electron (or hole) concentration gradient and random collision, **generation-recombination** of electron-hole pairs, **trapping-detrapping** or **capture-emission** of electron (or hole) at an electron (or hole) bound state of a defect or impurity, and **tunneling**. Drift and diffusion are described and modeled individually in the first two subsections, 31n and 32n, and then modeled together in subsection 33n by an effective potential or field, known as the quasi-Fermi potential or quasi-Fermi field. Generation-recombination and trapping involve creation and destruction of an electron and/or a hole. They are modeled initially in subsections 340 and 350 via phenomenological lifetimes in order to formulate the Shockley equations of transistor physics and engineering. The generation-recombination-trapping-tunneling (GRTT) kinetics of electrons and holes are described in subsections 36np. Their mathematical representations by the phenomenological electron and hole lifetimes are described in subsections 37n. The physical mechanisms and selected experimental data of the generation, recombination, trapping and tunneling rates of electrons and holes are given in subsections 38n.

Drift current dominates in MOS and junction-gate field-effect transistors (FETs). Diffusion current dominates in bipolar junction transistors (BJTs) and p/n junction diodes. Generation, recombination, trapping and tunneling cause leakage current in p/n junctions, low current gain in BJT, instability, aging, and failure. But GRTT are also the operating mechanisms that make some devices work.

310 Drift

The current carried by an electron travelling at a velocity v is - qv. The number of electrons in a volume Adx illustrated in Fig.310.1 is nAdx where n is the macroscopically averaged electron concentration in number/cm_3 (or number/m^3). If these electrons all move out of this volume in the +x direction during the time interval dt, then the electric current is

$$I_x = dQ/dt = -qnAdx/dt = -qnAv_x \tag{310.1}$$

where $v_x = dx/dt$ is the x-component of the electron velocity. The areal density of the electric current, known as **current density** in Ampere/cm_2 (or A/m_3), is then

$$J_x = I_x/A = -qnv_x. \tag{310.2}$$

We shall use the unit cm (centimeter) instead of m (meter) because it has been used in transistor physics since the invention of transistors.

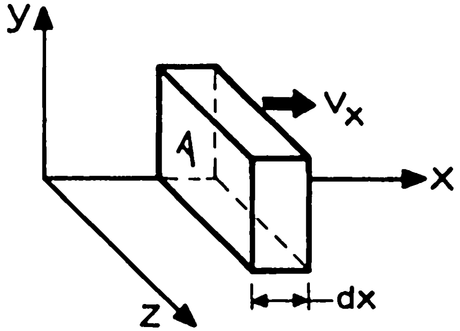


Fig.310.1 The coordinate system and the volume element used to calculate the drift current and drift current density of an electron.

311 Drift Velocity in an Electric Field

The electron velocity, v, can be calculated if the force acting on the electron is known. To illustrate, we compare the electron drift in semiconductor with that in vacuum or free space. In free space, electron obeys Newton's Law of motion, while in solids and semiconductors, it follows both Newton's Law and the **statistical law of randomness**. We consider these two cases in the presence of a spatially and temporally constant electric field in the x-direction, E_x.

In vacuum, Newton's Force Law for the electron is

$$m dv_x/dt = F_x = -qE_x. \qquad (311.1)$$

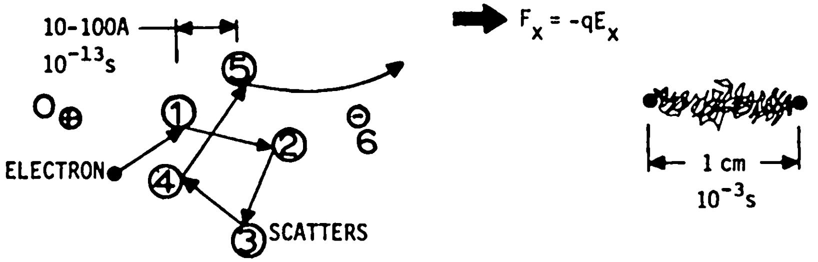
which is integrated to give the electron velocity

$$v_x = v_0 - (q/m)E_x t. \qquad (311.2)$$

Let the electrons start from rest, then $v_0=0$ and

$$v_x = (q/m)E_x t. \qquad (311.2A)$$

The velocity diverges if the duration of acceleration is large. The fallacy of infinitely large velocity arises from our implicit assumption that a constant electric field or constant force of acceleration exists over the infinitely large space making the constant force acting on the electron perpetually.



(a) Free Paths. (b) Drift Distance

Fig. 311.1 Random scattering of electron by scattering centers in a semiconductor under an applied electric field. (a) Expanded view of scattering centers (circles) representing the randomly vibrating host atomic cores and the few (1 part per million) randomly located impurities (dot). (b) Condensed view showing only a few of the 10^{11} scattering events. The electron trajectories are parabolic if the scatterers are not charged but are shown as straight lines for ease of drafting.

In a semiconductor, Newton's Law is modified by the forces from the **randomly vibrating** charged atomic cores or ions and by the **randomly located** impurity ions. They scatter the electrons via Coulomb's Law of electric force between charges. Consider the events shown in the expanded view given by Fig.311.1(a) and a hypothetical case in which the scatterers are neutral and are hard spheres. The electron is accelerated between two scattering centers labeled 1 and 2. During the transit between locations 1 and 2, there is no scattering force acting on the electron (the neutral scatterer assumption) so that the electron makes a **free flight** between 1 and 2, following Newton's Law and under the influence only of the externally **applied** electric force. At the end of the free flight, the electron velocity is given by v_x. The duration of the free flight is denoted by τ_f which is known as the **free time**. Then, integrating Newton's equation, Force = $-qE_x$ = $m dv_x/dt$ given by (311.1), the final electron velocity after the free flight is

$$v_x = v_0 - (q/m)E_x \tau_f. \qquad (311.3)$$

v_0 is the initial velocity before this free flight or the final velocity after the previous free flight. At the end of the free flight, the electron encounters another scatterer and experiences
another scattering event. Each scattering event changes the direction of the electron velocity and also its magnitude back to v_0 by absorbing the kinetic energy gained by the electron during its preceding free flight before encountering the hard-sphere scatterer. In a laboratory observation or measurement time interval, the electron will scattered about 10^{14} times in one second, as depicted by the unexpanded view in in Fig.311.1(b). Thus, we must take an average over all the scattering events to arrive at a theoretical expression of current in order to account for the experimentally measured current. This average over time, or **time average**, gives the **drift** velocity v_d.

$$\langle v_x\rangle = \langle v_0\rangle - (q/m)E_x\langle\tau_f\rangle \equiv -(q/m)E_x\tau \equiv v_d. \qquad (311.4)$$

The average, $<v_0>$, over all scattering events in the time interval of a measurement is zero because v_0 is completely random in direction due to the random vibration of the host atomic cores and random location of the impurity ions. The initial velocity, v_0, is also known as the **thermal velocity** since random vibration of the atomic cores is measured by the temperature of the sample. In the one-dimension example, randomness means that there are as many scattering events which terminate with a positive v_0 as there are with a negative v_0. Thus, their contributions to the average current observed during an infinitely long measurement period cancel out exactly. They cancel almost completely in a finite observation period. The incomplete cancellation is the origin of one type of noise known as **thermal noise**, resistor noise or Johnson noise. It is defined and measured by the root-mean-square average $v_{noise} = \sqrt{<[v_0-<v_0>]^2>} = \sqrt{<v_0^2>}$. The rms noise in the drift velocity is given by $v_d(\text{noise})=(q/m)E_x\sqrt{<\delta\tau_f)^2>}$ which is much smaller than the drift velocity $<v_x>$ and much more smaller than the thermal velocity noise just computed $\sqrt{<v_0^2>}$. Thus, the proposal of mobility fluctuation noise is fundamentally faulty. (See problems 311.1. and 312.10.)

In (311.4) the simplified symbol τ is used to denote the average free time $<\tau_f>$. It is the average of all the free times, one from each of the scattering events experienced by one electron. It is known as the **mean free time**.

The description just given for a hypothetical solid of an uncharged or neutral-hard sphere scatterer model can also be applied to the realistic situation. However, the algebra of the scattering mechanics is much more complex because the main scatterers in a real solid are the randomly vibrating and positively charged atomic cores. The scattering force between the charged atomic cores and electrons is coulombic, $1/r$, which acts at a distance and does not terminate as a hard sphere in the hypothetical model. But the conclusions are the same as our simplified model which has enabled us to demonstrate with the simplest mathematics. Some additional features are discussed next.

The descriptive, **drift**, comes about because the electron velocity due to random motion from random scattering or the thermal velocity, v_0, is much larger than the additional velocity gained by the electron from acceleration by the applied electric field. Thus, the electron drifts slowly in the direction of the applied force while it moves rapidly and randomly at a much higher velocity. For example, the thermal velocity of electrons at room temperature (300K) is about 10^7 cm/s which is computed from the average kinetic energy, $E_{ave}=(3/2)mv_0^2=(3/2)kT$ where k is the Boltzmann constant and T is the absolute temperature. The drift velocity can be estimated using (311.4). Assuming an electric field of 1 V/cm or 100 V/m, then the drift velocity is $(q/m)\tau E_x = (1.60\text{x}10^{-19}/9.11\text{x}10^{-31})10^{-13}\text{x}100 \approx 2$ m/s = 200 cm/s which is 5,000 smaller than the thermal velocity. Thus, there is a **slow drift superimposed** onto the **rapid random motion** of the electron when an electric field is applied. This is pictured in Figs.311.1(a) and (b).

From the above analysis, we immediately notice that if the electric field is 1 million times higher, 10^6 V/cm, the drift velocity would be $2\text{x}10^8$ cm/s or 200 times larger than the thermal velocity. Then thermal motion would not be important and the electron would accelerate continually without limit as if it were in a vacuum. This is untenable because as the electron drift velocity increases towards the thermal velocity, the electron kinetic energy also rises which increases the rate of electron energy loss to the vibrating atomic cores. The energy absorbed by the vibrating atoms is then propagated away to the surface and is dissipated by radiation or by collision of the surface atoms with the ambient gas molecules. This increasing rate of electron energy loss inside the solid will eventually equal the rate of energy gain by the electrons from acceleration by the applied force (dE_{loss}/dt = input power = applied force times velocity) to give and maintain a steady-state, known as macroscopic kinetic **power balance**, sometimes known as energy balance, a misnomer since it is the balance of input and output energies per unit time or power. At this steady state, the drift velocity is **saturated** to a value similar to the thermal velocity and the saturated value is known as **scattering-limited velocity** or **scatter-limited velocity**. This velocity saturation phenomenon at high applied electric field was first theorized successfully by Shockley in 1951 after many attempts by other solid state theorists during the preceding decade. Shockley coined it the **hot electron effect**. The term '**hot carrier effects**' is commonly used since the velocity saturation phenomenon also occurs for holes. Velocity saturation at high electric field is one of the several **hot electron** or **hot carrier effects** in high electric fields. Other hot carrier effects will be described in the following subsections on generation-recombination-trapping mechanisms. Hot carrier effects cause MOS and bipolar transistors to age and integrated circuits to fail. But, they are also the basic mechanisms of operation of some very useful Si devices and integrated circuits such as the UV EPROM (the Ultra-Violet Light Erasable Programmable Read Only Memory), the voltage reference p/n junction diode, and the avalanche p/i/n junction diode photon detector.

There are other less restrictive and more general ways of calculating the average than the simple model just described in which we traced the motion of one electron over a long time period. The other methods include all electrons at all energies or kinetic energies and are known as the **ensemble average**. The different ensemble methods are distinguished by the size of the element of the ensemble of elements. In the elementary ensemble average, its element is a group of specially designated or selected electrons, for example, all the electrons in a volume element dxdydz at a time interval dt. One can readily see that the argument made for the zero average value of v_0 in (311,4) also applies to the ensemble average. Ensemble average is the main theme in Statistical Mechanics, a specialty in physics. It develops the mathematics of the many ensemble averages of different element sizes, such as the canonical ensemble, microcanonical ensemble, and the grand canonical ensemble; and it connects the mathematical results to experimentally measurable quantities such as the temperature, electrical and heat conductivity, and others.

As we recall, time average is taken over all the scattering events experienced by one electron in a volume element dxdydz during a time interval dt. In contrast, the ensemble average is taken over all the electrons in a volume element dxdydz at an instance of time, t. Frequently, the ensemble average of a parameter (such as scattering rate) is taken over all the particles of a species such as all the electrons in the conduction band or all the holes in the valence band, for example, to get the mobilities, the reciprocal scattering rate or the free time is averaged over all the electrons at the various energy levels or kinetic energies and velocities in the valence and conduction bands. Presumably ensemble average gives the same result as the time average if the operation variable or the material parameter is 'sufficiently' averaged so that noise or random fluctuation is smoothed out.

The simple time average model just described by tracing one electron has a fundamental limitation. That is, the measurement or observation time of the average current, dt, must be large compared with the mean free time but also all the free times so that the electron is scattered many times. Otherwise, the average is not stationary and will have another value in a later observation time interval, t_1 to t_1+dt_1. If the measured values at different times are very different, then there is a large amount of noise. The ensemble average also has a fundamental limitation. The volume element, dxdydz, must be sufficiently large in order to contain many electrons. Otherwise, the averages would have large fluctuations from one location (dxdydz) to another ($dx_1dy_1dz_1$) in a uniform sample.

Ensemble and time averages can be combined to give a quick answer. Reliability engineers have taken advantage of such an approach. For example, a hundred integrated circuit chips from one production run can be tested at a range of high electric fields and high temperatures for a relatively short time (such as a few days) to give statistically significant operating life data. This data then enables the engineer to calculate the aging rate and extrapolate the projected operating life (such as 10 years) of the integrated circuit from this and similar production runs.

312 Drift Current, Drift Mobility and Conductivity

The results just obtained in vacuum and in a solid can be substituted into the current density equation, (310.2) to give

$$J_x = -qnv_x = qn(qt/m)E_x \quad \text{(Vacuum)} \tag{312.1A}$$

$$= qn(q\tau/m)E_x \quad \text{(Solids, Semiconductors).} \tag{312.1B}$$

An **electron mobility** can be defined as the average velocity per unit electric field,

$$\mu = |v_x/E_x| \equiv q\tau/m \tag{312.2}$$

Then, the current density, (312.1B), can be written as

$$J_x = qn\mu E_x \tag{312.3}$$

where

$$\mu_n = qt/m \quad \text{(Vacuum)} \tag{312.4A}$$

$$\mu_n = q\tau/m \quad \text{(Semiconductors, Solids).} \tag{312.4B}$$

It is evident from the above results that the electron mobility increases continuously with time in vacuum, while it is a temporal constant in a semiconductor. The difference comes from the fact that there is no scattering in vacuum while there are many random scattering events in a semiconductor or solid. These random scattering events change the electron (or hole) velocity randomly and the electron has only a small net drift motion in the direction of the applied force as indicated in Figs.311.1(a) and (b).

The electron drift produces a net or nonvanishing current known as the **drift current** while in contrast, the random electron motion produces no net current since the random component of its velocity averages zero. The drift current can also be written as a product of the conductivity of the solid and the electric field,

$$J_x = qn\mu_n E_x = \sigma_n E_x \tag{312.5}$$

where

$$\sigma_n = qn\mu_n . \tag{312.6}$$

is the **conductivity** due to electrons in the conduction band, abbreviated as the **electron conductivity**.

A similar drift-current equation can be obtained for holes. The pair of drift current density equations in the x-direction and their generalization to three dimensional are

$$J_{nx} = q\mu_n nE_x \qquad J_n = q\mu_n n\mathbf{E} = \sigma_n \mathbf{E} \tag{312.7A}$$

and

$$J_{px} = q\mu_p pE_x \qquad J_p = q\mu_p p\mathbf{E} = \sigma_p \mathbf{E}. \tag{312.7B}$$

313 Temperature Dependences of the Drift Mobility

The drift velocity is a measure of how much scattering the electron or hole experiences in a semiconductor. A larger amount of scattering would result in a smaller drift mobility. If there are two different types of scattering events, then the reciprocal free time would be added and then averaged, since the reciprocal free time is proportional to the probability of scattering. Instead of adding the probability or reciprocal free time before average, a good approximation is to add after average which then allow us to add the reciprocal of the mobility of each of the scattering mechanisms. For two scattering mechanisms, we can use the approximations

$$\frac{1}{\tau} = \frac{1}{\tau_1} + \frac{1}{\tau_2} \tag{313.1}$$

$$\frac{1}{\mu} \approx \frac{1}{\mu_1} + \frac{1}{\mu_2} \tag{313.2A}$$

and

$$\frac{1}{\sigma} \approx \frac{1}{\sigma_1} + \frac{1}{\sigma_2} = \frac{1}{q\mu_1 n_1} + \frac{1}{q\mu_2 n_2} \tag{313.2B}$$

where the combined or measured conductivity can be used to define a combined mobility, μ, given by $\sigma = q\mu(n_1+n_2)$ if there are two distinct groups of electrons. This is known as the **Matthiessen's rule**. The approximation is accurate enough for diodes and transistor applications.

In solids, there are two fundamental scatterers which determine scattering free time and mean free time of electrons and holes. They limit the carrier mobility and conductivity. One is **intrinsic** and **dynamic**, arising from the random vibration of the host atoms or the host ion cores. In addition to maximum conductivity of a regular conductor, it also determines the condition of infinite conductivity or superconductivity or the condition when **randomness ceases** or **coherency holds**. The second is **extrinsic** and **static**, arising from the randomly located physical defects and foreign impurities, such as the atomic disruption at the solid surface or interfacial boundary layer, and the artificially added or residual chemical impurities.

They share a common force origin that causes the electron orbit to bend, the electrostatic Coulomb force. One significant difference, at least in the mathematical analysis of scattering, is the dynamic versus the static scatterer. The concept of phonon or quantized lattice vibration waves is introduced to ease the algebra and to use the classical particle collision concepts to describe the dynamic scatterers. Although the impurity ions also vibrate, just like the hosts but at different frequencies, it is not the vibration but (i) the excess or deficient charge of the impurity ion compared with the host and (ii) the random location of the impurity that are responsible for random scattering of the electrons by the impurity ions. The key common feature is randomness. Scattering by the vibrating atomic cores or ions is known as **thermal scattering**, phonon scattering or lattice scattering. Scattering by the impurity ions is known as impurity scattering.

Lattice scattering can be divided into acoustic phonon scattering, optical phonon scattering, polar phonon scattering and intervalley phonon scattering due to the complex band structures of semiconductors. Phonons are quanta of energy from quantization of the lattice vibration waves to be illustrated shortly. Impurity scattering can also be subdivided into ionized and neutral impurity scattering. The free time can be calculated using the quantum mechanical transition rate theory for an electron making a scattering transition from one state or energy level to another state or energy level. The mean free time can be calculated by the method of statistical mechanics to average the free times. These are given in advanced courses. However, the approximate formulae predicting the temperature dependence of the rate of these scattering mechanisms can be derived based on simple physical reasoning using elementary mathematics. These are discussed in the following three subsections, one of which introduces the concept of phonons or quantized lattice vibration waves.

Ionized Impurity Scattering

For ionized impurity scattering, the scattering probability is proportional to the final density of state of the scattered electron since higher final density of state gives a larger number of states for the electron to be scattered into. The final density of state has an energy dependence of $\sqrt{E}$ [from D(E) of (233.5) using the reference $E_C=0$]. It is also proportional to the effective area of the Coulomb Potential well of the impurity ion seen by the incident electron, since larger effective scattering area gives higher probability of scattering. This effective area is a measure of the range of the ion potential seen by the electron and commonly known as the **scattering cross section**. The effective area may be estimated by a circle whose radius r is inversely proportional to the electron energy as suggested by the Bohr model given in section 141 where the energy is given by (141.8), $E_n \propto n^{-2}$, and the radius is given by (141.6), $r_n \propto n^2$, so that the radius is inversely proportional to the energy, $r_n \propto E_n^{-1}$. Thus, the effective scattering area is inversely proportional to the square of the electron energy. This E^{-2} dependence is also given by the well-known Rutherford scattering formulae on atomic scattering of an electron by a helium ion. Thus, the total scattering probability of an electron by an impurity ion, which is inversely proportional to the mean free time of scattering, is then proportional to

$$\begin{aligned}\frac{1}{\tau} &\propto \text{(Final Density of State)} \times \text{(Scattering Area)}\\ &\propto \sqrt{E} \times E^{-2} = E^{-3/2} \qquad (313.3)\end{aligned}$$

The mobility of the electrons in the conduction band is then obtained by averaging over all electrons with all possible energies in the conduction band weighed by its number at each energy. The number of electrons at each energy level is proportional to the Fermi-Dirac distribution function or the occupation factor. For low density, the F-D distribution reduces to the Boltzmann distribution or the exponential factor, exp(-E/kT). Combining these three factors (the scattering rate, the density of state, and the fraction of occupied states), and performing the average over all the electrons in the conduction band, $E=E_C$ to $E_{C'}$, we obtain an analytical expression for the mobility. Using the mobility definition obtained in (312.4B), $\mu = q\tau/m$, we can then factor out the temperature variable in the mobility formulae of ionized impurity scattering,

$$\mu_I = (e/m)\ \langle\tau_I\rangle \propto (e/m)\langle E^{3/2}\rangle \propto (kT)^{3/2}. \tag{313.4}$$

Thus, the temperature dependence of the mobility due to impurity ion scattering is

$$\mu_I = A_I T^{3/2}. \tag{313.5}$$

This is a physically reasonable result because at higher temperatures, there are more electrons with higher kinetic energy or moving at higher speed. Thus, their orbit or path would be bent less by the Coulomb force exerted on them by the randomly located impurity ions. Less random changes of trajectory means less random scattering and hence higher mobility which is predicted by (313.5). However, the power law dependence on temperature given above can only be arrived at by considering all three factors which determines the average scattering rate of the electrons by the impurity ions just discussed. The prefactor, A_I, is inversely proportional to the concentration of the impurity ion, N_I, since higher concentration means more scatterer, more scattering events, or lower mobility. Thus,

$$A_I = A_{I0}/N_I$$

and

$$\mu_I = A_{I0}(T^{3/2}/N_I) \tag{313.6}$$

Phonons for Describing Lattice Scattering

Scattering of electrons (and holes) by the randomly vibrating charged atomic cores due to Coulomb force can be most easily described by the classical particle collision picture by quantizing the lattice vibration waves known as phonons. Thus, we shall first describe the dynamics of atomic vibration on a lattice, the quantization of the vibrational waves, and the frequency-wavenumber (ω_q-q) or energy-momentum (E_q-$\hbar q$) relationships of the lattice vibration waves. The lattice vibration wavenumber will be denoted by q to contrast it with electron wavenumber k. The description of ω_q-q is necessary since there are many phonons that can scatter an electron (or a hole) and each gives a different energy and momentum exchange.

The one-dimensional lattice vibration waves and the ω_q-q diagrams are shown side-by-side in Figs.313.1 and 313.2 which we shall explain.

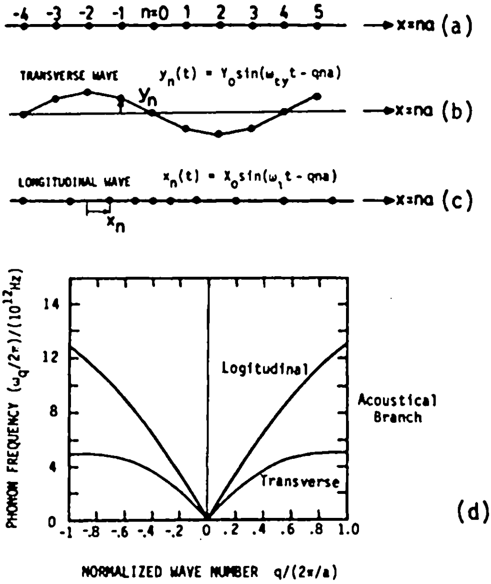

-4 -3 -2 -1 n=0 1 2 3 4 5
x=na (a)
TRANSVERSE WAVE
$y_n(t) = Y_0\sin(\omega_{ty}t - qna)$
y_n
x=na (b)
LONGITUDINAL WAVE
$x_n(t) = X_0\sin(\omega_l t - qna)$
x_n
x=na (c)
PHONON FREQUENCY $(\omega_q/2\pi)/(10^{12}Hz)$
0
4
8
12
14
Longitudinal
Transverse
Acoustical Branch
-1 -.8 -.6 -.4 -.2 0 .2 .4 .6 .8 1.0
NORMALIZED WAVE NUMBER q/(2π/a)
(d)

Fig.313.1

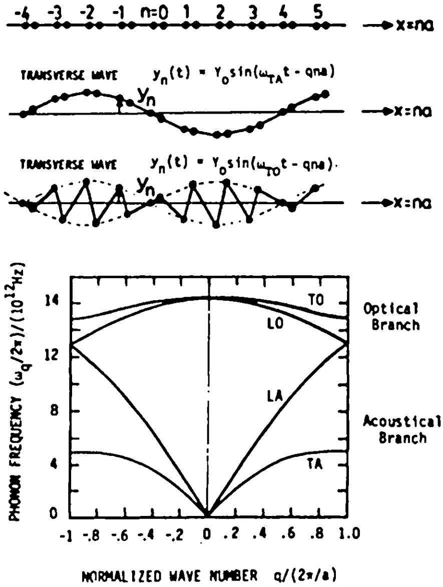

-4 -3 -2 -1 n=0 1 2 3 4 5
x=na
TRANSVERSE WAVE
$y_n(t) = Y_0\sin(\omega_{TA}t - qna)$
y_n
x=na
TRANSVERSE WAVE
$y_n(t) = Y_0\sin(\omega_{TO}t - qna)$
y_n
x=na
PHONON FREQUENCY $(\omega_q/2\pi)/(10^{12}Hz)$
0
4
8
12
14
TO
LO
LA
TA
Optical Branch
Acoustical Branch
-1 -.8 -.6 -.4 -.2 0 .2 .4 .6 .8 1.0
NORMALIZED WAVE NUMBER q/(2π/a)

Fig.313.2

Fig.313.1 Phonons in a one-dimensional monoatomic lattice. (a) Equilibrium. (b) Transverse acoustical wave, λ=8a. (c) Longitudinal acoustical wave, λ=8a. (d) Lattice vibration or phonon spectra, acoustical branch.

Fig.313.2 Phonons in a one-dimensional diatomic lattice. (a) Equilibrium. (b) Transverse acoustical wave, λ=8a. (c) Transverse optical wave, λ=8a. (d) Lattice vibration or phonon spectra, acoustical and optical branches.

Figure 313.1(a) shows a monoatomic one-dimensional lattice or one-dimensional chain of atoms at equilibrium. It is also the configuration at the instance of time when each atom (or atomic core) is located at its equilibrium position. The x-axis is chosen to parallel the length direction of the one-dimensional lattice or chain. Figure 313.1(b) shows the displacement of the atomic cores from their respective equilibrium position when a transverse wave of atomic vibration is propagating along the chain of atoms or the x-direction. It is evident that the wavelength is 8a. The wave number is then $q=2\pi/\lambda=2\pi/8a=\pi/4a$. Figure 313.1(c) shows the displacements when a longitudinal wave is propagating along the x-direction with also a wavelength of 8a or wave number of $q=2\pi/\lambda=2\pi/8a=\pi/4a$.

These displacement waves can be represented by $A\cos(\omega_q t\text{-}qX_{n0})$, $A\sin(\omega_q t\text{-}qX_{n0})$, or the real or imaginary part of $A\exp[i(\omega_q t\text{-}qX_{n0})]$. ω_q is the angular frequency. q is the wave number given by $q=2\pi/\lambda$. X_{n0} is the position of n-th lattice point or n-th equilibrium atomic position, which can be replaced by $X_{n0}=na$ where n = ... -3, -2, -1, 0, +1, +2, +3,...

For the longitudinal displacement wave, the amplitude of the displacement of each atom is $A=\mathbf{i}_x X_0$ and the frequency is $\omega_q=\omega_l$. Denote the **longitudinal displacement of the n-th atom** from its equilibrium position by $x_n(t)$ as shown in Fig.313.1(c). It is then mathematically given by

$$\begin{aligned} x_n(t) &= X_n(t) - X_{n0} = X_n(t) - na \\ &= X_0\cos(\omega_l t - qna) = X_0\cos(\omega_l t - n\pi/4) \end{aligned}$$

where $q=2\pi/\lambda=2\pi/8a$ for Fig.313.1(c). This longitudinal displacement wave can also be represented by the sine or complex exponential form.

For the two transverse displacement waves, $A=\mathbf{i}_y Y_0$ and $\omega_q=\omega_{ty}$ or $\mathbf{i}_z Z_0$ and $\omega_q=\omega_{tz}$. Similar equations for the transverse displacements of the n-th atom, $y_n(t)$ and $z_n(t)$, can also be written down, such as, $y_n(t)=Y_0\sin(\omega_{ty}t-qna)$.

Because the atomic cores are point masses, the frequency-wave number relationship (ω_q-q) is not linear or the velocity of the wave $v=d\omega_q/dq$ is not a constant (like the electromagnetic waves in vacuum) but is a function of ω_q or q. This is known as dispersive. Furthermore, because the force between neighboring atomic cores (electrostatic or Coulomb force) are generally dependent on the direction of displacement, the three vibration frequencies in the three displacement directions are generally not equal, $\omega_l \neq \omega_{ty} \neq \omega_{tz}$.

The wave number dependences of the vibration frequency, ω_q-q, are shown in Fig.313.1(d) for the one-dimensional monoatomic lattice with the assumption $\omega_{ty}=\omega_{tz}=\omega_t$. It illustrates that the longitudinal wave generally vibrates at a higher frequency than the transverse wave of equal wavelength or wave number.

These lattice vibrational waves can be quantized in accordance with the Planck and de Broglie hypotheses. The quantum of energy is $E_q=\hbar\omega_q$ and the wavelength is $\lambda_q=2\pi/q=h/p$ where p is the momentum and $p=\hbar q$. This quanta is known as the phonon. The scattering of the electrons by these lattice vibrational waves (due to the Coulomb electrostatic force between point charges) can be analyzed classically as collisions between electrons and phonons. The laws of conservation of energy and momentum are strictly adhered to. If the reduced k-space, $-\pi/a \leq k < +\pi/a$, is used, then a slight extension of de Broglie momentum definition is made due to the presence of the crystal's periodic potential, $p=\hbar(k+2m\pi/a)$, where m is a positive or negative integer including zero. In addition, phonons can be created to dissipate electron energy and destroyed to increase electron energy during inelastic electron-phonon collisions.

Each value of ω_q at a given q is known as a phonon mode. The electronic counterpart of the phonon mode is the **electron state** described by the E-k (energy-wavenumber) electronic energy band diagram developed in chapter 1.

Audible and ultrasonic sound waves propagating in solid are compressional or longitudinal waves. Thus, these two ω_q-q curves are known as the **longitudinal and transverse acoustical branches** of the lattice vibration spectra. The phonons are called the **longitudinal** and **transverse acoustical phonons**.

The interatomic distance, a, is about five angstroms (10^{-8}cm). It is much smaller than the wavelength of a 1000Hz audio sound wave (about 0.35m in air, 1.5m in water, 5m in Al, 2m in Au, 8m in Si, and 6m in SiO_2). Thus, the wave number q of the sound wave propagating in a solid or semiconductor is very small and is near the origin, or $q \simeq 0$ in Fig.313.1(d). This is the origin of the term 'acoustical branch'. Sound wave in solid is a longitudinal compressional-expansional wave whose speed is $d\omega_q/dq$ of the longitudinal acoustic wave at $q=0$.

In Si, GaAs and other binary compound semiconductors, there are two atoms in a primitive cell. The second atom allows the atoms to vibrate at two frequencies at a given wavelength along either the longitudinal or transverse direction. This is illustrated in Figs.313.2(a)-(c) for transverse displacement. Figure (a) shows the equilibrium configuration of a one-dimensional chain with two atoms per cell. Figure (b) shows the lower-frequency transverse acoustical mode at a wavelength $\lambda=8a$. Figure (c) shows the higher-frequency transverse mode at the same wavelength 8a in which the adjacent atoms displace in the opposite directions. Figure (d) shows the frequency-wavenumber ω_q-q or dispersion diagram. The higher frequency modes are known as the optical branches, optical phonons, or **transverse and longitudinal optical phonons**. The term 'optical' originates from the fact that these higher frequencies fall in the infrared range of the optical spectra and the optical phonons interact strongly with the infrared photons.

In a real crystal, the lattice vibration spectra (ω_q-q diagram) is three-dimensional because the lattice vibration waves propagate in all three space directions. Figures 313.3(a)-(f) give the experimental-theoretical spectra of five face-centered cubic crystals (Si, Ge, GaAs, ZnSe, and ZnS) and the hexagonal SiO_2 (α-quartz). Figures 313.3(a') and (e') show the q-space (same as k-space) direction and point labels. The experimental data were obtained from thermal neutron diffraction measurements and fitted to a multi-parameter but physics-based theory.

Consider Si first. Figure (a) gives the ω_q vs q along the three principle symmetry directions, $<q_x,0,0>$, $<0,q_y=q_z>$ and $<q_x=q_y=q_z>$, or $<1,0,0>$, $<1,1,0>$ and $<1,1,1>$ directions. The longitudinal modes are labeled L and the transverse modes are labeled T. The q-axis normalization is $2\pi/a$ where $a=5.43A$ is the lattice constant of Si.

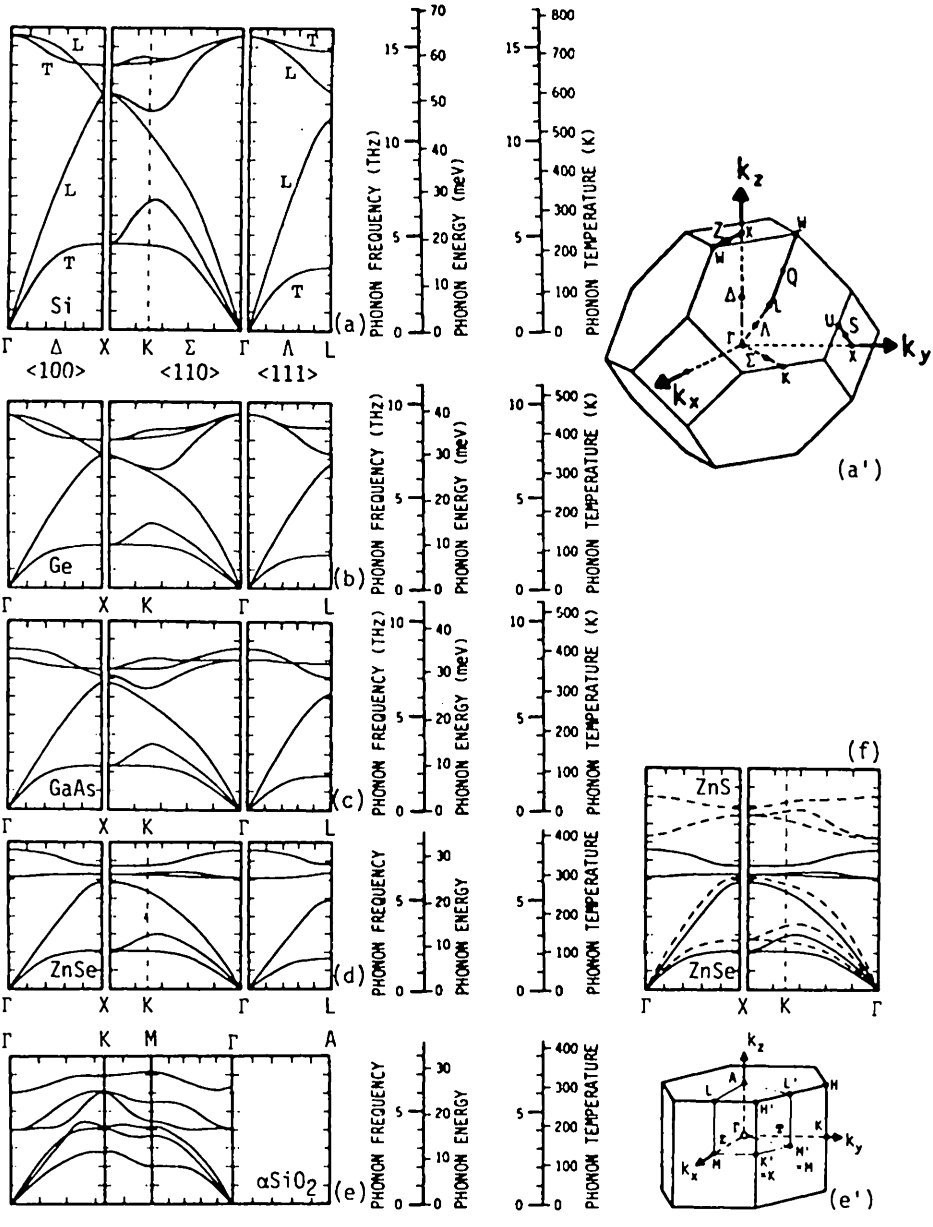


Fig.313.3 Experimental-theoretical phonon spectra of (a) Si, (b) Ge, (c) GaAs, (d) ZnSe, (f) ZnS and ZnSe; and (e) SiO_2 α-quartz. First Brillouin zone of (a') face-centered cubic and (e') hexagonal lattices.

Figure 313.3(b) shows the phonon spectra of Ge which have lower vibration frequencies than Si because Ge atom is heavier than Si atom and so oscillates slower. Figure 313.3(c) shows GaAs' transverse acoustical branch is almost identical to Ge because $M_{Ga}+M_{As}=69.72+74.92\approx 2M_{Ge}=2x72.59$. However, the M_{As}-M_{Ga} mass difference removes the LO-TO(Γ), LA-LO(X) and LA-LO(L) degeneracies in the Ge spectra to give the frequency gaps in GaAs. Figures (b)Ge, (c)GaAs, and (d)ZnSe show the trend in an isoelectronic sequence due to increasing mass difference and interatomic force or ionicity: frequencies are lower and gaps are larger in ZnSe because $M_{Se}-M_{Zn}>M_{As}-M_{Ga}>M_{Ge}-M_{Ge}=0$. Figure (f) compares ZnSe with ZnS. ZnS's optical branch oscillates at higher frequencies due to lighter sulfur than selenium $M_S<M_{Se}$ and ZnS's gaps are larger due to larger mass difference, $M_{Zn}-M_S=65.38-32.06=33$ $\gg$ $M_{Se}-M_{Zn}=78.96-65.38=14$. Figure 313.3(e) gives the phonon spectra in the basel plane of the hexagonal SiO_2 (α-quartz). The three branches come from the three SiO_2 molecules (9-atoms) in the primitive hexagonal unit cell. Additional higher vibration frequency branches would be expected due to Si and O vibrations not as an SiO_2 unit.

Two scales are added to the frequency axis of the phonon spectra: the energy scale $E_q=\hbar\omega_q$ in milli-electron-volt (meV) and an equivalent temperature scale defined by $k_BT_q=\hbar\omega_q$. They are especially useful to help understand the physics of electron-phonon interaction such as scattering of the electrons by the vibrating atoms. The energy scale, $E_q=\hbar\omega_q$, and the wavenumber or momentum scale, $p=\hbar q$, can be used to visualize energy and momentum conservation, as well as phonon generation or emission (atom set to oscillate by bouncing electrons) and phonon destruction (absorption by the electrons) during the electron-phonon collisions. The effective phonon temperature, T_q, gives an indication of the number of phonons with energy E_q or frequency ω_q which are available or excited at a given lattice temperature, T_L, since the number of phonons is given by the Bose-Einstein function, $n_q=1/[\exp(\hbar\omega_q/k_BT_L)-1]=1/[\exp(T_q/T_L)-1]$. In addition, the energy or **frequency density of phonon modes** (analog of the density of electron states) is an important factor that determines the rate of electron scattering by phonons. This mode density is given by the same expression as that of the electron density of state in chapter 2, with the electron wave vector **k** replaced by the phonon wave vector **q**. Thus, the **phonon density of mode** is $dq_xdq_ydq_z/(2\pi)^3$. The factor of two due to electron spin and Pauli's exclusion principle (fermion) is absent for phonon because phonon is spinless (boson).

Lattice Scattering

The rate of scattering electrons by phonons and the phonon-scattering-limited mobilities have a temperature dependence different from that of ionized impurity scattering which will now be described. Due to spatial orthogonality, the electron drift velocity is affected only by the longitudinal (parallel to the applied force) phonons and not by the transverse phonons. Consider the long wavelength acoustic phonon first. Since it is essentially at $q=0$ and $\omega_q=0$, thus the scattering is elastic.

Like that of the static impurity scatterer, it neither creates nor destroys any phonons. Thus, the phonon number is constant and need not be considered which simplifies the analysis. For the same reason as that given for scattering by impurity ions, the probability of electron scattering by phonons is proportional to the final electron density of state which gives a $\sqrt{E}$ dependence. The probability is also proportional to the square of the lattice vibration amplitude or the cross sectional area of the vibrating silicon (or host atom) cores. The reason was again similar to that of the isolated impurity ion except for two differences: (i) the silicon atoms in phonon scattering are distributed over the entire crystal while only one isolated impurity ion was calculated and then the contributions from all the impurity ions were added by means of $A_I \propto N_I^{-1}$, and (ii) the assumed impurity ion is fixed while the host atoms (Si) are randomly vibrating about their equilibrium positions. Although the impurity ions do vibrate just as the host atoms, the basis for (ii) is that scattering due to the vibration of the impurity ions is already included in (i) while impurity scattering comes from the random location of the excess or deficient impurity ion charge, $(Z_{ion} - Z_{host})q$. The random vibration of the impurity atoms will increase impurity scattering slightly which has been neglected by theorists.

The lattice temperature or $(3/2)k_BT_L$ measures the average kinetic energy of the vibrating Si or host atomic core, but the kinetic energy is proportional to the square of the vibration or displacement amplitude, which is in turn proportional to the effective scattering cross section. Thus, the scattering rate of electrons by the randomly vibrating Si (or host) atoms is proportional to

$$\begin{aligned} \tau_A^{-1} &\propto \text{(final density of state)}\cdot\text{(scattering area)} \\ &\propto \sqrt{E}\cdot(kT). \end{aligned} \tag{313.7}$$

Averaging over all the electrons in the conduction band using the procedure for scattering by the impurity ions, then the electron mobility limited by acoustic phonon scattering is

$$\mu_A = (e/m)\langle\tau_A\rangle \propto \langle 1/(kT\sqrt{E})\rangle \propto (kT)^{-3/2}. \tag{313.8}$$

This gives a temperature dependence of

$$\mu_A = A_A T^{-3/2}. \tag{313.9}$$

Next, consider the more complex inelastic phonon scattering mechanisms in which a phonon is either created (generated) or destroyed (absorbed) by the electron (or hole) during scattering. These are: (i) electron scattering to another energy band minimum (or valley) of the six conduction band minima of Si or GaP, known as intervalley phonon, (ii) hole scattering by an optical phonon, in nonpolar semiconductors (C, Si, and Ge); and (iii) electron and hole scattering by a longitudinal polar phonon in polar semiconductors (GaAs, InP, and others) and in insulators (NaCl, KF, SiO_2 and Si_3N_4). The main difference between the elastic acoustic phonon and these inelastic phonons is that electrons will interact or be scattered by these inelastic phonons at a well defined phonon energy to satisfy both energy and momentum conservation. These inelastic phonon energies are one to two k_BT_L/q or 20 to 60 meV (as indicated in Fig.313.3) which are comparable to the average kinetic energy of the electron. In contrast, the acoustic phonon energies are negligible compared with the average electron energy in elastic acoustic phonon scattering. Since few high energy phonons (high means about $2k_BT$ or 50meV) are excited at room temperature, 300K, the phonon-absorption events are less probable than the phonon-emission events during scattering of an electron. In the latter scattering event, a phonon of energy $\hbar\omega_O \approx 50$meV is emitted (generated) from the scattering site. But the Energy Conservation Law requires that the electron kinetic energy be greater than $\hbar\omega_O$ in order to create the phonon. Thus, the probability of electron scattering via phonon emission is reduced by the factor $\exp(-\hbar\omega_O/kT)$ to account for the fewer electrons or lower electron densities at the higher kinetic energies, $E_k > E_q = \hbar\omega_q$. Again taking into account the final density of state of the electrons, the probability of inelastic scattering of an electron via optical or intervalley phonon emission is then

$$\tau_O^{-1} \propto \sqrt{E}\cdot\exp(-\hbar\omega_O/kT). \qquad (313.10)$$

Thus, the temperature dependence of the mobility is

$$\mu_O = A_O T^{-1/2}\exp(\hbar\omega_O/kT). \qquad (313.11)$$

The mobility formula becomes an integral if scattering due to absorption of an optical or intervalley phonon is included. This integral cannot be evaluated analytically. The complete formula is very complex and is derived using quantum statistical mechanics in advanced graduate courses. But from simple physics, we have extracted the key mechanism, viz, phonon emission. The result, (313.11), is physics based, covers a wider range of temperature, and is more accurate than the widely used empirical power law, T^n, in textbooks, engineering articles, and Si integrated circuit design handbooks.

The combined mobility of all three scattering processes can then be approximated by the Matthiessen rule i.e.

$$1/\mu = 1/\mu_I + 1/\mu_A + 1/\mu_O. \qquad (313.12)$$

The experimental mobilities in Si are shown in Figs.313.4 and 313.5. Figure 313.4 shows the temperature dependences of the electron and hole mobilities in pure Si. Note the $T^{-3/2}$ dependence which agrees with theory, (313.9), of longitudinal acoustical phonon scattering. Figure 313.5 shows the impurity concentration dependences of the majority carrier mobilities at 300K, electron mobility in n-Si and hole mobility in p-Si. The data have also been fitted to empirical formulae similar to those we just derived. These are tabulated in Table 313.1. The more accurate inelastic phonon scattering formulae, (313.11), was replaced by an unphysical and less accurate power law in this empirical fit.

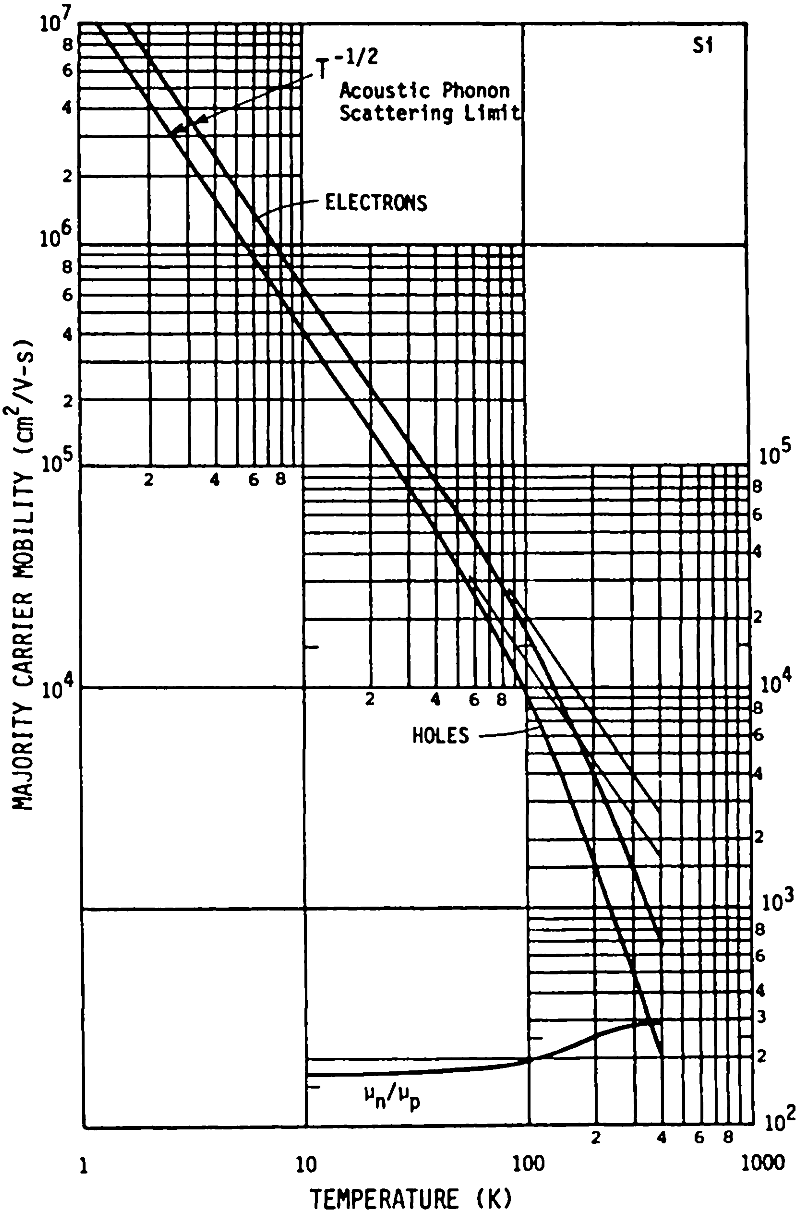


Fig.313.4 Intrinsic mobility of electrons and holes in pure Si due to phonon scattering as a function of temperature.

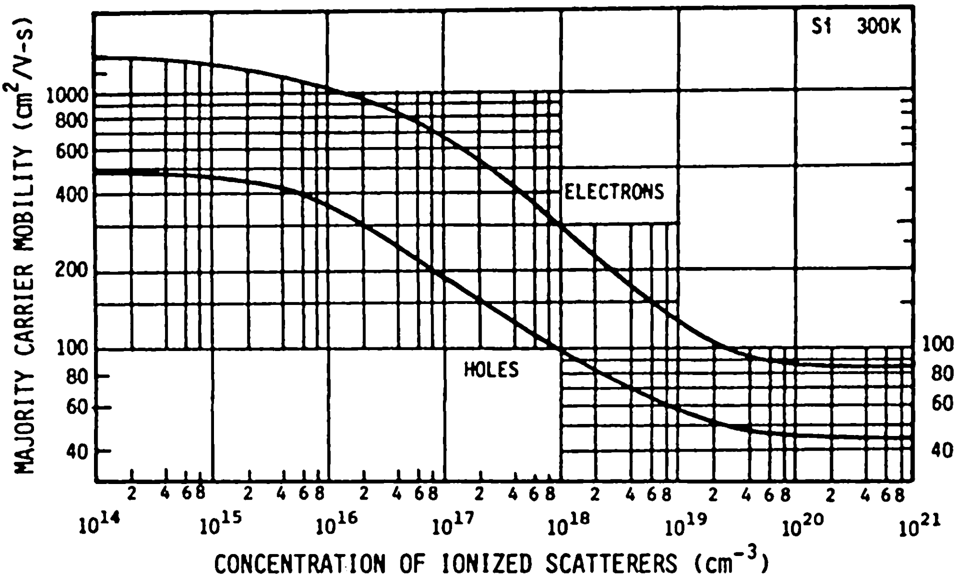


Fig.313.5 Majority carrier conductivity mobilities in impure Si vs ionized impurity scatterer density at 300K.

Table 313.1
Empirical Parameters of Carrier Mobilities in Silicon ($cm^2/V\text{-}s$)

	ELECTRONS	HOLES
μ_A	$2.18\text{x}10^7T^{-1.5}$	$1.30\text{x}10^7T^{-1.5}$
μ_0	$1.22\text{x}10^{11}T^{-3.13}$	$6.64\text{x}10^{10}T^{-3.25}$
μ_I	$= \mu_\infty + \mu_{I0}(N_0/N_I)^{\alpha}(T/300)^{\beta}$	
μ_∞	90	45
μ_{I0}	To be determined by students as a problem.	
N_0	$1.3\text{x}10^{17}$	$6.3\text{x}10^{16}$
α	0.91	0.72
$\beta\alpha$	1.5	1.5

314 Electric Field Dependence of Mobility

As the electric field increases, the electrons are accelerated and their drift velocity increases. When the electron kinetic energy exceeds the optical phonon energy, optical phonons are generated. The generation rate is so large that few electrons could exceed this energy. Thus, a drift velocity limit, θ_n, is reached which can be estimated from the energy balance equation, $(1/2)m_n\theta_n^2=\hbar\omega_O$, giving

$$\theta_{Ln} = \sqrt{2\hbar\omega_O/m_n} \approx 10^7 \text{ cm/s}. \qquad (314.1)$$

The numerical value assumes $\hbar\omega_O=50\text{meV}$ and $m_n=m$. It is roughly equal to the thermal or rms velocity of electrons at room temperature. A similar derivation can be made for holes.

This is a high electric field or hot electron effect. It was first theorized and analyzed by Shockley in 1951. [W. Shockley, "Hot electrons in germanium crystal and Ohm's law," Bell System Technical Journal, 30(10), 990-1034, October 1951] The maximum velocity is known as the **scattering limited velocity** or **saturation velocity** and its effects on transistors are known as **velocity saturation effects**.

The drift velocity at low electric field is proportional to the electric field. This constancy was used to define a low field mobility, μ_0, in (312.2), giving $v_d(E\rightarrow0)=\mu_0E$. The initial rise with increasing electric field and the final saturation to a constant at high fields suggest the following empirical approximation for the electric field dependence

$$v_d(E) = \mu_0E/[1 + (E/E_c)] = \theta_L/[1 + (\theta_L/\mu_0E)]. \qquad (314.2)$$

The electric field dependent mobility can be approximated using $\mu=v_d/E$

$$\mu(E) = \mu_0/[1 + (\mu_0E/\theta_L)^\gamma]. \qquad (314.3)$$

An empirical exponent, γ, is added in (314.3) to fit the data better. An accurate formulae was first derived by Shockley in 1951 in which the energy and momentum conservation laws are applied to electron-phonon scattering and phonon emission and absorption are also included. Figure 314.1 gives the experimental data of the electron and hole drift velocities in pure silicon at 300K. The theoretical curves were computed by this author in 1969 by numerically integrating the mobility integral. They were based on Shockley's simple theory but included an optical phonon and an acoustical phonon for both scattering (momentum) and energy exchange. The theory accounts for the velocity maximum, known as velocity overshoot recently, and the negative differential conductance. Acoustical phonon scattering and acoustical phonon emission account for the further rise of the drift velocity when the field exceeds about 10^6 V/cm because optical phonon emission cannot keep up with the increasing energy gain by the electrons at the higher electric field. The fundamental cause of the velocity maximum and negative differential conductance was understood in 1951 and 1969 but seems to be still mysterious to recent authors who performed theoretical calculations using the empirical Monte Carlo method instead of the physics-based analytical method of Shockley. [For example, see Fig.5 of Physical Review B41(17), 12122-12128, 15 June 1990.]

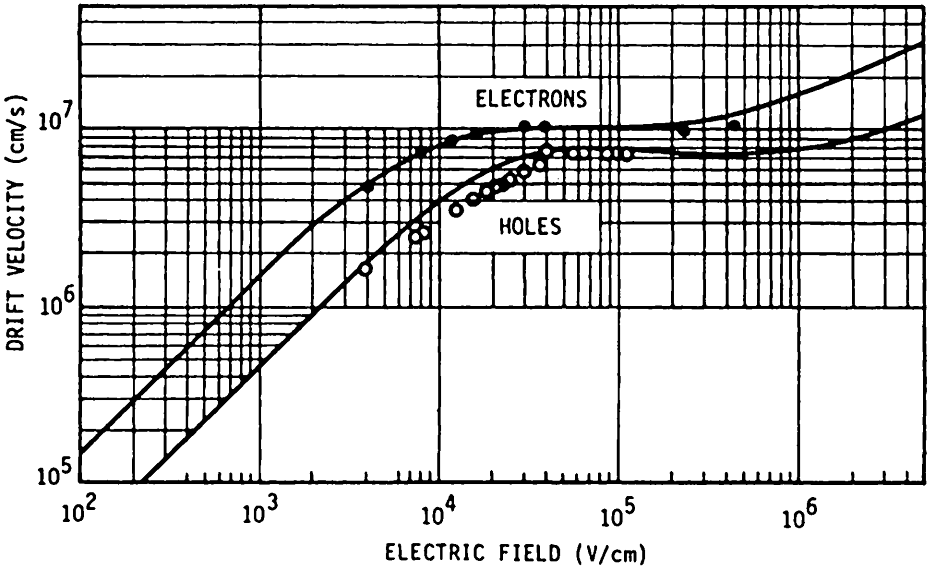


Fig.314.1 The electric field dependences of the drift velocity of electrons and holes in silicon at room temperature (300K). The data were from the published literature and the theory was computed by this author in 1969 using Shockley's 1951 model containing two phonons, an acoustical phonon and an optical phonon, and both phonons conserve momentum (scatter) and energy (electron energy loss or dissipation). E_c=420V/cm (electrons). E_c=1100V/cm (holes).

315 Intrinsic and Extrinsic Conductivities of a Semiconductor

The electrical conductivities of pure and impure homogeneous semiconductors are now derived using the foregoing results. Numerical examples are then given.

The total drift current in a semiconductor in an applied electric field is the sum of the electron and hole drift currents. Let the electric field be in the x-direction. Then the total drift current per unit area or the drift current density, obtained by using (312.5) and (312.6), is

$$J_x = J_{nx} + J_{px} = \sigma_n E_x + \sigma_p E_x = (\sigma_n + \sigma_p)E_x = \sigma E_x. \qquad (315.1)$$

This shows that the total conductivity of a semiconductor, σ, is the sum of the conductivities of electrons, σ_n, and holes, σ_p, defined in (312.6A) and (312.6B) respectively,

$$\sigma = \sigma_n + \sigma_p = q\mu_n n + q\mu_p p. \tag{315.2}$$

Intrinsic Conductivity of Pure Semiconductor

In a pure semiconductor, the electron and hole concentrations are equal and given by the intrinsic carrier concentration, $N=P=n_i$. The conductivity in a pure sample is known as the **intrinsic conductivity** and its reciprocal is known as the **intrinsic resistivity**. The intrinsic conductivity is given by

$$\sigma_i = \sigma_{ni} + \sigma_{pi} = q\mu_n n_i + q\mu_p n_i = q(\mu_n + \mu_p)n_i. \tag{315.3}$$

For silicon at 300°K (room temperature), the intrinsic carrier concentration is about 10^{10} cm^{-3}. More precisely, at 296.5K, $n_i=1.00\times10^{10}$ cm^{-3} and at 300K, $n_i=1.332\times10^{10}$ cm^{-3} while a less accurate value of 1.45×10^{10} cm^{-3} is given by some books. In a pure Si at 300K, the electron and hole mobilities are due to lattice scattering and they have the values of $\mu_n=1422$ cm^2/V-s and $\mu_p=478$ cm^2/V-s. In verbal discussions and in this elementary textbook, the rounded figures of 1500 and 500 cm^2/V-s as well as 10^{10} cm^{-3} are frequently used. Using the accurate values, the intrinsic conductivity and resistivity of silicon at 296.57K are

$$\begin{aligned}\sigma_i &= q(\mu_n + \mu_p)n_i \\ &= 1.602\times10^{-19}(1422+478)\times1.00\times10^{10} = 3.04\times10^{-6} \text{ mho-cm},\end{aligned} \tag{315.4}$$

and

$$\rho_i = 1/\sigma_i = 1/3.04\times10^{-6} = 329\text{k ohm-cm}. \quad (1\text{k}=1000\neq1024) \tag{315.5}$$

Conductivity of Impure Semiconductors

When the semiconductor is homogeneously doped with an impurity, the electron and hole concentrations at thermal equilibrium will differ from n_i in order to maintain detailed balance, $NP=n_i^2$, and charge neutrality, $\rho=q(P-N+N_{AA}-N_{DD})$. For an n-type semiconductor doped with a donor impurity, the electron concentration will increase above its intrinsic value and the hole concentration will decrease below its intrinsic value. It takes very little impurity doping to cause such a change. (Here, 'little' means in the range of parts per billion of host atom whose concentration is of the order of 10^{22-23} host-atom/cm^3.) For example, one part per billion of impurity atom in silicon amounts to $10^{-9}\times5\times10^{22} = 5\times10^{13}$ impurity atoms/cm^3 where 5×10^{22} atoms/cm^3 is the Si atomic density. Even at such a low concentration of phosphorus donor impurity in silicon, the electron and hole concentrations are drastically changed from its intrinsic value of 10^{10} cm^{-3}. At 296.57K, (242.8A) and (242.9A) give

$$N \simeq N_{DD} = 5\times10^{13}\text{cm}^{-3} \quad >> n_i = 10^{10}\text{cm}^{-3} \tag{315.6}$$

and

$$P = n_i^2/N \simeq 10^{20}/5\times10^{13} = 2\times10^{6}\text{cm}^{-3} << n_i = 10^{10}\text{cm}^{-3}. \tag{315.7}$$

The electron (or majority carrier) and hole (or minority carrier) conductivities are:

$$\sigma_n = q\mu_n N = 1.6\times10^{-19}\times1422\times5\times10^{13}$$

$$= 1.14\times10^{-2}\text{S/cm} = 1/(8.80\Omega\text{-cm}) \tag{315.8}$$

and

$$\sigma_p = q\mu_p P = 1.6\times10^{-19}\times 478\times2\times10^{6}$$

$$= 1.52\times10^{-10}\text{S/cm} = 1/(6.54\times10^{9}\Omega\text{-cm}). \tag{315.9}$$

Thus, the minority carrier conductivity can be neglected compared with the majority carrier conductivity and the total conductivity is nearly equal to the majority carrier conductivity. This is known as the **extrinsic conductivity** since the most carriers come from a foreign or extrinsic source, i.e., from the phosphorus donor impurity atoms which dope the silicon crystal to n-type in the above example. Note that the extrinsic conductivity, 1.14×10^{-2}S/cm given by (315.8), is much larger (3000 times larger) than the intrinsic conductivity, 3.04×10^{-6}S/cm given by (315.4).

320 DIFFUSION

Electrons and holes move continuously and randomly in a semiconductor because they are scattered by the randomly vibrating ionic cores of the host atoms and by the randomly located impurity ions. The microscopic force is the electrostatic 1/r Coulomb Law between two point charges. In solids, the total force acting on one electron is the sum of the 1/r Coulomb force from the many ions, which are not static but vibrating, and from the many other electrons, which are distributed and some randomly moving in the crystal. The electron (or hole) concentration is a macroscopic quantity averaged over a sufficiently large volume element dxdydz containing many electrons so that the microscopic variations are averaged out. Nevertheless, it may not be macroscopically constant from one volume element dxdydz to another, $dx_1dy_1dz_1$. For instance, it becomes spatially varying when an external force is applied to the semiconductor sample at a localized spot, such as a temperature gradient in the ambient gas whose molecules are bombarding the surface, or a noncontact force acting a distance via an electric, magnetic or electromagnetic force or photon. The carrier concentration can also be spatially varying if the concentration of the dopant impurity ions varies spatially. For example, each group-V donor ion would produce one electron, giving a spatially dependent electron concentration if the donor concentration varies spatially.

When there is a spatial variation of the electron (or hole) concentration, electrons (or holes) will move from the location of higher concentration to the location of lower concentration. This is known as **diffusion**. Diffusion occurs because more electrons (or holes) are scattered in the volume of higher electron (or hole) concentration, say dxdydz, than in the volume of lower electron (or hole) concentration, say $dx_1dy_1dz_1$. Thus, there are more electrons scattered out of the volume of higher electron concentration, dxdydz, into the lower-concentration surrounding, $dx_1dy_1dz_1$, than there are electrons scattered in the reverse direction. This net flow of electrons from locations of higher concentration to locations of lower concentration is known as diffusive flow or diffusion.

The diffusive flow of electrons ceases when equilibrium or homogenization is reached, that is, when the electron and hole concentrations become individually constant in space. This will occur in a homogeneous sample, where N_{DD}=constant and N_{AA}=constant. But, diffusive flow will never cease when $N_{DD}=N_{DD}(x)$ or $N_{AA}=N_{AA}(x)$. Instead, in an inhomogeneously doped sample, an internal electric field is built up which produces an electron and hole drift current. Equilibrium or electronic equilibrium can and will be reached inside inhomogeneously impurity doped samples when the net electron and net hole currents are each zero. This is the condition of electron drift current exactly balancing the electron diffusion current, and simultaneously, the hole drift current exactly balancing the hole diffusion current. The spatial variation of the impurity ion density will produce a nonvanishing current due to ion diffusion and drift. However, these ion currents are so small at room temperatures that they can be neglected for all practical purposes so that ionic equilibrium can be assumed at room temperatures. Thus, the foregoing partial equilibrium consists of two approximate equilibria: (i) the electrical (electronic and ionic) equilibrium defined by zero net particle flux or current of each particle species (electrons, holes and ions in this case), and (ii) the thermal equilibrium of the electrons with the phonons defined by the condition of zero net energy flux flowing between any two macroscopic ensembles of particles. One may have electrical equilibrium or zero current but not exact thermal equilibrium, such as the open circuit condition of a solar cell which is close to thermal equilibrium. One may also have electrical (electronic or ionic or both) nonequilibrium at nearly thermal equilibrium, such as a low test current passing through a solid to measure its conductivity.

In section 313, we have discussed the important scattering mechanisms which determine the magnitude and the temperature variation of the electron and hole mobilities. These were the acoustical phonon scattering mechanism, the optical or intervalley phonon scattering mechanism, and the ionized impurity scattering mechanism. These very same scattering mechanisms also determine the magnitude and the temperature dependences of the diffusion rate of electrons and holes. The parameter which measures and characterizes the diffusion rate is the **diffusion coefficient** or **diffusivity**. They have also been known as the diffusion constant, a misnomer, since generally they are not constant and can vary with position, temperature, electric field and even with time. The diffusivity is denoted by the symbol D_n for electrons and D_p for holes. Since diffusion and drift are both controlled by the same scattering mechanisms, we would expect the diffusivity and the mobility to be closely related to each other. Indeed they are proportional to each other through the **Einstein Relationship** as we shall demonstrate by elementary mathematics in the following paragraphs.

The flux of electrons and holes is proportional to their concentration gradient and the proportionality constant or coefficient is the diffusion coefficient, the diffusion constant or the **diffusivity**. Since the particles flow away from a region of higher concentration, the particle flux is proportional to the **negative** of the concentration gradient, -dn/dx for electrons and -dp/dx for holes in one dimension. Multiplied by the electron charge or the hole charge (with sign), we get the expression for the **areal density** of the diffusion current (A/cm^2)

$$J_{NX} = (-q)(-D_n dn/dx) \quad \text{(1-d or one-dimensional)} \tag{320.1A}$$

$$J_N = +\, qD_n \nabla n \quad \text{(3-d or three-dimensional)} \tag{320.1B}$$

$$J_{PX} = (+q)(-D_p dp/dx) \quad \text{(1-d or one-dimensional)} \tag{320.2A}$$

$$J_P = -\, qD_p \nabla p \quad \text{(3-d or three-dimensional)} \tag{320.2B}$$

where D_n and D_p are the diffusivity of electrons and holes respectively.

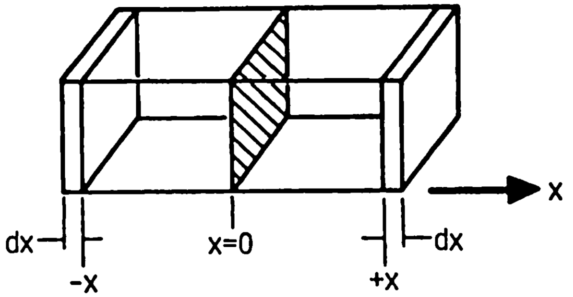


Fig.320.1 The geometry and coordinate used to derive the one-dimensional diffusion current.

A macroscopic model can be developed to relate the diffusivity to the **scattering probability** or the **scattering mean free time**. Such a model was used in sections 310-312 to describe the drift current and drift mobility of the electrons and holes in an electric field. Consider the one-dimensional case shown Fig.310.1, which is expanded in Fig.320.1. In this one-dimensional case, all the electrons are restricted to move only in the +x or -x directions. (It is actually more general: it is the x-component of the general three-dimensional physical situation.) The particle flux is the number of particles passing through a unit area per second. To calculate the electron flux passing through the y-z plane or x=0 plane in Fig.320.1, we need to determine the fraction of electrons located in the half spaces on the right and left of the x=0 plane and we need to write down the number of electrons which pass through the x=0 plane from both half spaces because of scattering in the right and left half spaces. Consider those electrons initially in the slice located between the x and x+dx planes. The probability that these electrons will cross the x=0 plane is giving by the following factors:

(1) The probability that they are scattered out of dx is dx/λ.
(2) The probability that they are scattered out of dx in the -x direction to cross the $x=0$ plane is 1/2 since half of the electrons will move in the +x direction.
(3) The probability that they are not scattered in the region between the planes $x=0$ and x so that they will reach the plane $x=0$ is $\exp(-x/\lambda)$.

The parameter λ is the **free path** or the distance travelled by the electron between successive scattering events. Obviously, it is related to the **free time** of flight, τ_f, introduced previously during the derivation of the drift current and drift mobility. The relationship is $\lambda = v_o\tau_f$ where v_o is the **random initial velocity** of the electron starting each free flight. Thus, the total electron particle flux crossing the $x=0$ plane in the negative x-direction is

$$F_{\leftarrow} = \frac{1}{2}\int_0^{\infty} n(x)v_x \frac{dx}{\lambda}\exp(-x/\lambda) \tag{320.3}$$

$$= \frac{1}{2}\int_0^{\infty}\left[n(0) + x(dn/dx)_0\right]v_x \frac{dx}{\lambda}\exp(-x/\lambda) \tag{320.3A}$$

$$= \frac{1}{2}\left[n(0)v_x + \frac{dn(0)}{dx}v_x\lambda\right]. \tag{320.4}$$

Here, v_x is the electron velocity restricted to flow in the x-direction and $(dn/dx)_0 \equiv dn(0)/dx$ is the $dn(x)/dt$ evaluated at $x=0$. In the approximation leading to (320.3A), we expanded n(x) in Taylor series and retained terms only to the first order, $n(x) = n(0) + (x^1/1!)[dn(0)/dx] + (x^2/2!)[d^2n(0)/dx^2] + \ldots \approx n(0) + (x/1!)[dn(0)/dx]$.

A similar calculation can also be made for the electrons crossing the $x=0$ plane in the +x direction from the left half of semiconductor located in $x= -\infty$ to 0. It is given by

$$F_{\rightarrow} = \frac{1}{2}\int_{-\infty}^{0} n(x)V_x \frac{dx}{\lambda}\exp(+x/\lambda) \tag{320.5}$$

$$= \frac{1}{2}\int_{-\infty}^{0}\left[n(0) + x(dn/dx)_0\right]v_x \frac{dx}{\lambda}\exp(-x/\lambda) \tag{320.5A}$$

$$= \frac{1}{2}\left[n(0)v_x - \frac{dn(0)}{dx}v_x\lambda\right]. \tag{320.6}$$

The net electron flux crossing the x=0 plane in the +x direction is then

$$F_n = F_{\rightarrow} - F_{\leftarrow} = - v_x\lambda \frac{dn}{dx} = - D_n \frac{dn}{dx} . \qquad (320.7)$$

D_n is the diffusivity and is related to the free path and free time by

$$D_n = v_x\lambda = v_x{}^2\tau \qquad (320.7A)$$

where we have used $\lambda = v_x\tau$ and τ is the free time (same as τ_f used before) between successive scattering events. Since the electron velocity is distributed from 0 to ∞, the measured diffusivity in experiments is an average over all the electrons whose velocities extend the entire range. This average is usually written as

$$D_n = \langle v_x\lambda \rangle = \langle v_x{}^2\tau \rangle . \qquad (320.8)$$

The symbol, $\langle \ \ \rangle$, denotes an average. Two average algorithms can be used: the average over many scattering events, known as the time or **temporal average**, and the average over many electrons each undergoing a scattering event, known as the **ensemble average**. The drift mobility was computed using the time average model in section 311. A similar derivation gives the hole particle flux:

$$F_p = - D_p \frac{dp}{dx} . \qquad (320.9)$$

Equations (320.7) and (320.8) are known as the **Fisk's law**. The 1-d electrical currents, (320.1A) and (320.2A), due to electron and hole diffusion are obtained from multiplying the Fisk equation for the particle flux by the particle charge. For the three-dimensional case, the above derivation procedure still applies and the same results are obtained for each dimension. For example, the definition for the electron diffusivity given by (320.8) remains applicable but the average is made over a three-dimensional distribution of electron velocity or electron kinetic energy. A modification is necessary if the diffusivity and mobility are directional dependent which occurs in anisotropic crystals such as the piezo-electric crystals and high-temperature superconductors. In which case, the diffusivity and mobility are represented by 3x3 matrices or tensors.

The total electron and hole currents can now be obtained by adding the diffusion current to the drift current obtained in (312.7A) and (312.7B). These are:

$$J_N = q\mu_n nE + qD_n\nabla n \qquad J_{Nx} = q\mu_n nE_x + qD_n(\partial n/\partial x) \qquad (320.10)$$

$$J_P = q\mu_p pE - qD_p\nabla p \qquad J_{Px} = q\mu_p pE_x - qD_p(\partial p/\partial x) \qquad (320.11)$$

We have not imposed any restrictions on the variables and parameters other than that they are macroscopically averaged over many particles in a volume element $\Delta x\Delta y\Delta z$ and many collision events in a time interval Δt. Thus, the parameters and variables can be both space and time dependent. Their space-time dependences can be explicitly written out, for example, $\mu_n=\mu_n(x,y,z,t)$, $n=n(x,y,z,t)$, $E=E(x,y,z,t)$, $J_N=J_N(x,y,z,t)$. It is evident that a notation convention must be used to help to remember the physical meanings of symbols because of the large number of variables and parameters. The IEEE convention of symbols is used in this book, thus, the upper case J used for the total current variable is replaced by the lower case, $j_N(x,y,z,t)$ while the upper case represents the average or d.c. value. The IEEE convention will be defined shortly when the Shockley Equations are presented. The definition is described in Appendix A at the end of this book.

321 The Einstein Relationship

The diffusivity and the mobility are related by the Einstein Relationship

$$D/\mu = (kT/q). \tag{321.0}$$

This can be proved as follows. The mobility is given by the average defined by $\mu=(e/m)<E\tau>/<E>$ where the average < > is the ensemble average over all the electrons (holes) in the entire energy band. For a function F, it is defined by

$$< F > = \int_0^\infty F \exp(-E/kT)\sqrt{E}dE + \int_0^\infty \exp(-E/kT)\sqrt{E}dE. \tag{321.1}$$

$\sqrt{E}$ comes from the density of electronic state in the bands. exp(-E/kT) comes from the low density or Boltzmann approximation of the Fermi-Dirac distribution function, $f(E) = 1/\{1+\exp(E-E_F)/kT]\} = \exp[-(E-E_F)/kT]$, which gives the fraction of the states occupied by electrons at low electron concentrations. Using $E=mv^2/2$, $<E>=3kT/2$, and $<v_x^2\tau>=<v_y^2\tau>=<v_z^2\tau>=<v^2\tau>/3$, then from D_n given by (320.8) and μ_n given by the ensemble average of (312.4B) which is $\mu_n=(q/m)<\tau E>/<E>$, we have

$$D_n/\mu_n = <v_x^2\tau>[(q/m)<E\tau>/<E>]^{-1} = kT/q. \tag{321.2}$$

A similar derivation can be carried out for holes to give

$$D_p/\mu_p = kT/q. \tag{321.3}$$

322 The Boltzmann Relationship

An alternative derivation of the Einstein Relationship or a derivation of the Boltzmann Relationship at thermal equilibrium can be made by balancing the diffusion and the drift current since the total electron (and hole) current is zero at equilibrium by the definition of equilibrium or thermal and electrical equilibria. The derivation of the Einstein Relationship is attained using the Boltzmann Relationship between the electron (or hole) concentration and the electric potential. One can readily carry the algebra in the reverse direction to derive the Boltzmann Relationship assuming the Einstein Relationship is known since the Einstein

Relationship can be independently derived from kinetic theory and statistical mechanics which led to (321.2) and (321.3).

We carry out the derivation of the Einstein Relationship. At thermal and electrical equilibrium, the net current of each particle species is zero. Thus, j_{Nx} (or $\mathbf{j}_N$ in 3-d) given by (320.10) is zero. We use $E_x = -dV/dx$ (or $\mathbf{E} = -\nabla V$ in 3-d) where V is the electrostatic or electric potential at the position x. The electron concentration is given by (233.8) at low electron concentration when the Boltzmann approximation applies. E_C is the reference of the potential energy for electrons since the kinetic energy of electrons with total energy E is $E-E_C$. The basis of using E_C as the potential energy is the same as that given for the shallow donor impurity bound state in paragraph (4) of section 223, i.e., the periodic potential (whose reference energy is the vacuum level) is already contained in the effective mass of the quasi-particle, the electron. This can be rigorous proved by deriving the effective mass equation from the many-electron Schrödinger equation which is often given in an advanced course. Thus, we can write $E_C = -qV$ and the electron concentration is then given by

$$N(x) = N_C\exp[-(E_C-E_F)/kT] = N_C\exp(E_F/kT)\exp(qV/kT) \qquad (322.1)$$

Using this in the electron current density given by (320.10) and setting it to zero since the electron current is zero at equilibrium, then

$$\begin{aligned} 0 = J_{Nx} &= q\mu_n N E_x + qD_n(\partial N/\partial x) \\ &= q\mu_n N(-\partial V/\partial x) + qD_n N(\partial V/\partial x)(q/kT) \\ &= -qN(\partial V/\partial x)[\mu_n - qD_n/kT]. \end{aligned} \qquad (322.2)$$

Thus,

$$D_n/\mu_n = kT/q. \qquad (322.3)$$

One can readily generalize this derivation to the three-dimensional space.

The preceding analysis may be viewed as an alternative proof or derivation of the Boltzmann relation between the carrier concentration and the electric potential, if we accept the Einstein relationship as the starting point. This is admissible since the Einstein relationship can be proved directly using kinetic theory and statistical mechanics such as (321.2) and (321.3).

323 Examples of Diffusion Current

We shall not give idealized examples of diffusion current for hypothetical situations which are traditionally included in many textbooks. Instead, we focus on realistic examples, i.e., the useful transistor devices. These application examples are given in chapters 5 and 7 on p/n junctions and bipolar junction transistors. Diffusion of photo-generated minority carriers in photoconductors and solar cells are additional examples of diffusion currents.

330 CONSTANCY OF THE FERMI ENERGY LEVEL

The Fermi energy or Fermi level, E_F, was first defined in the derivation of the thermal equilibrium distribution function of electrons and holes in section 231. It is independent of position, i.e. spatially constant, and it does not change with time, i.e., time invariant or stationary. **It is a complete constant in space-time and it is meaningful and defined only at thermal equilibrium.** Suppose that it were to vary with position or time, then, the macroscopically averaged electron and the hole concentrations are neither spatially constant nor stationary with time, and the sample is no longer at equilibrium.

To demonstrate physically and mathematically that E_F is independent of position, we suppose that it is not independent of position. Thus, if we replace the electron concentration, n, in the electron current density equation, (320.10), by the Boltzmann approximation, (322.1), $n(x,y,z,t) = N(x,y,z) = N_C\exp[-(E_C-E_F)/kT]$, then using $E_C=-qV$ and $E_x=-dV/dx=+dE_C/qdx$, we have

$$J_{Nx} = q\mu_n N E_x + qD_n(\partial N/\partial x) = -\mu_n N(dE_F/dx). \qquad (330.1)$$

But at thermal and electrical equilibria, the electron current must vanish (the hole current also vanishes). Since $N\neq 0$, we must have $dE_F/dx=0$ or E_F = spatially constant in order that $J_{Nx}=0$. A similar proof can be given to show the E_F is stationary or independent of time since $\partial J_{Nx}/\partial t = 0$.

A more general proof of the spatial constancy of the Fermi level can be given for a device structure with a heterogeneous material composition or two different materials in contact at an interface. We use the most drastic inhomogeneity, an abrupt change (step function) in the electrostatic potential or the conduction or the valence band edge shown in Fig.330.1. We suppose that E_F might have a step at the interface. The proof to follow will show that there is no step in E_F at the interface even when there is a step in E_C.

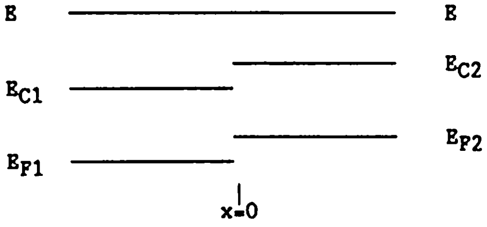


Fig.330.1 Energy band diagram with a potential step at x=0 to demonstrate the spatial constancy of the Fermi level.

To obtain the proof, we again use the fact that at equilibrium, the net electron (or hole) current flowing across the interface must vanish. Net means the inclusion of all components of current flowing in all directions. The proof proceeds as follows. The currents flowing to the right and left through the potential discontinuity at x=0 must be balance at any electron energy E. The current is given by integrating the electron flux over k_y and k_z at a given k_x since there is a wide distribution of the electron velocity components in the y and z direction at a given electron energy and velocity v_x or k_x. Thus, we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v_{x1} f_1 dk_{x1} dk_{y1} dk_{z1} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v_{x2} f_2 dk_{x2} dk_{y2} dk_{z2} \qquad (330.2)$$

where $f_1 = 1/[1+\exp(E-E_{F1})/kT]$, $f_2 = 1/[1+\exp(E-E_{F2})/kT]$, $v_{x1} = \partial E/\hbar\partial k_{x1}$ and $v_{x2} = \partial E/\hbar\partial k_{x2}$, the latter two are the group velocities discussed in section 190 for electron wavepackets which are impinging onto the potential step. The total energy is a constant since there is no energy gain or loss mechanisms, thus, $E = (k_{x1}^2+k_{y1}^2+k_{z1}^2)/2m_1 = (k_{x2}^2+k_{y2}^2+k_{z2}^2)/2m_2$. Substituting these into (330.2), and performing the integration over k_{y1}, k_{z1} on the left and k_{y2}, k_{z2} on the right, it is easily shown that

$$E_{F1} = E_{F2}. \qquad (330.3)$$

Thus, there cannot be a step between E_{F1} and E_{F2} as postulated in Fig.330.1.

This proof shows that the Fermi energy or Fermi level is a constant across an interface where there is an abrupt change of the electron or hole potential energy. Practical examples are the solid heterojunction interfaces between metal, semiconductor, and insulator, such as the two interfaces of the MOS (metal/oxide/Silicon) structure and the heterostructure junction Ge_xSi_{1-x}/Si. Liquid, gas, and solid interfaces are other examples. This proof can be readily extended to any location of a semiconductor where the material or potential energy change is not so abrupt.

331 THE QUASI-FERMI LEVELS AND QUASI-FERMI POTENTIALS

The result given by (330.1) indicates that if the current does not vanish, then the gradient of the Fermi level must not be zero, i.e., if $J_{Nx} \neq 0$, then $dE_F/dx \neq 0$. This suggests that we can use (330.1) to define a Fermi level when there is current flow. This definition is not unique but it is the most expedient since it reduces to the correct asymptotic result when the current approaches and becomes zero. In order not to be confused with the term 'Fermi level' which is defined and meaningful only at thermal and electrical equilibria, Shockley introduced the term quasi-Fermi level or imref (Fermi spelled backward). It is denoted by the symbol

F_N so that (330.1) can be used at nonequilibrium when $J_N \neq 0$. Thus, the conduction current defines the quasi-Fermi level. In one dimension, it is

$$J_{Nx} = \mu_n N dF_N/dx \tag{331.1}$$

$$= -q\mu_n N dV_N/dx \tag{331.1A}$$

and in three dimension, it is

$$J_N = \mu_n N \nabla F_N \tag{331.2}$$

$$= -q\mu_n N \nabla V_N \tag{331.2A}$$

where $F_N = -qV$ is the **quasi** or **nonequilibrium Fermi level** for electrons. V_N is the **quasi-Fermi potential** for electrons. The term for electrons must not be dropped since the electron current is generally different from the hole current under a nonequilibrium condition. This crucial difference between different species of particles (electrons, holes, trapped electrons, migrating ions, etc.) was overlooked prior to 1958 by early theorists of irreversible thermodynamics and nonequilibrium statistical mechanics who were developing nonequilibrium recombination theories via a small-signal or linear expansion at equilibrium and made a tactical error by assuming a single quasi-Fermi level for all species of particles. A similar definition can be made for holes using the hole current equation. In one dimension it is

$$J_{Px} = \mu_p P dF_P/dx \tag{331.3}$$

$$= -q\mu_p P dV_P/dx \tag{331.3A}$$

and in three-dimensional space, it is

$$J_P = \mu_p P \nabla F_P \tag{331.4}$$

$$= -q\mu_p P \nabla V_P. \tag{331.4A}$$

where $F_P = -qV_P$ is the **quasi** or **nonequilibrium Fermi level for holes**. V_P is the **quasi-Fermi potential for holes**. The negative sign arises since our energy scale is for electrons while the potential scale is for a positive charge.

Similarly, the nonequilibrium concentrations of electrons and holes can be expressed by or can be used to define their quasi-Fermi levels or potentials. Replacing E_F by F_N in the equilibrium electron concentration expressions (233.8) and (242.3) and similarly, replacing E_F by F_P in the equilibrium hole concentration expressions (233.10) and (242.4), we have

$$N = N_C \exp[-(E_C - F_N)/kT] \tag{331.5}$$

$$= n_i \exp(-(E_I - F_N)/kT] \tag{331.6}$$

$$= n_i \exp[q(V_I - V_N)/kT] \tag{331.7}$$

and

$$P = N_V \exp[-(F_P - E_V)/kT] \tag{331.8}$$

$$= n_i \exp[-(F_P - E_I)/kT] \tag{331.9}$$

$$= n_i \exp[q(V_P - V_I)/kT] \tag{331.10}$$

Thus, there are two key features or key interpretations for the quasi-Fermi levels and quasi-Fermi potentials. Their physics and physical bases are clearly evident. These are explained in the following two paragraphs.

Equations (331.1) and (331.2) show that the electron current is the electron conductivity, $q\mu_n N$, times the negative gradient of the quasi-potential for electrons. It is not the negative gradient of the electric potential, $-\nabla V_I$, which is the electric field given in the electron drift current expression, (312.7A), and which only accounts for the drift current and not the diffusion current. The reason is that part of the electron current is due to electron diffusion. Thus, the gradient of the quasi-Fermi potential for electrons, ∇V_N, takes into account both the diffusion and the drift current components of the total electron conduction current.

The second key feature, shown by (331.5), (331.6) and (331.7), is that the quasi-Fermi level and potential for electrons is just an alternative way of expressing the electron concentration at a general non-equilibrium condition. It is an **exponential transform** which transforms the electron concentration variable, n(r,t) or N(r), into two variables, the electric potential, $v_I(r,t)$ or $V_I(r)$, and the quasi-Fermi potential for electrons, $v_N(r,t)$ or $V_N(r)$. The choice of this transformation is logical and expected since it generalizes and extends the equilibrium Boltzmann relationship between the local particle concentration and the local potential to an arbitrary electrical nonequilibrium or nonzero current condition. Any other transformation can be used but none as mathematically convenient and physically appealing to help in remembering the physics.

These interpretations can also be repeated for the quasi-Fermi level and potential for holes. Under an arbitrary electrical nonequilibrium condition, one cannot expect the electron concentration and electron current to be related to those of holes by a simple universal relationship, such as $NP=n_i^2$. This dictates the use of a quasi-Fermi level and potential for holes, as indicated in (331.8)-(331.10) which are different from those for electrons, and still another different quasi-Fermi level and potential for a third particle species such as the electrons trapped at a substitutional impurity species or a mobile ion particle species. Thus, each species of particles requires a quasi-Fermi level and potential to represent its nonequilibrium concentration and its current.

There is indeed a deeper operational meaning and a deep fundamental concept underlying the quasi-Fermi levels and potentials. Beside the meaning 'not real' used in 'quasi-particles', the word 'quasi' means that even though particle and electrical currents exists, the energy distribution of the particle number or particle density, is still very nearly the same as the thermal equilibrium distribution which is the Fermi distribution for electrons or its Boltzmann approximation at low density, or the Bose distribution or its Boltzmann approximation for the phonons. Thus, quasi means nearly thermal equilibrium.

The fundamental reason of being able to have such a small deviation from the thermal equilibrium distribution is because the electrons are randomly scattered at a very high rate, 10^{14} scattering events in one second, by the randomly vibrating host atoms in the solid. It is the very frequent scattering at the tremendous rate that keeps the energy distribution (i.e. kinetic energy distribution or velocity distribution) at nearly the thermal equilibrium distribution. The distribution would become very different from the Fermi or Boltzmann distributions if the forces applied to the electrons, such as an electric field that gives drift current, is very high. An example is the hot electron effect on drift velocity which was discussed in section 314. In this case of a high applied electric field, the quasi-Fermi level and quasi-Fermi potential distribution function can still be used to represent the electrons. It is then purely an exponential transform to change the variable from N to V_I and V_N for mathematical expediency. But (331.5) to (331.7) may not be in the simplest forms. An improvement is to use an effective or electron temperature, T_e, which is different from the lattice or phonon temperature, T or T_L, that appeared in (331.5) to (331.7). This effective temperature can be defined by the average electron kinetic energy, $(3/2)k_BT_e=(\text{electron KE})_{ave}=<(3/2)m_ev^2>$. It would then account for the higher average kinetic energy of the electrons when they are accelerated by a high electric field.

Finally, there is indeed a deep exoteric classical concept underlying the quasi-Fermi potential and quasi-Fermi energy. To chemists, electrochemists, chemical engineers, and modern materials scientists, the quasi-Fermi potential for electrons is known as the **electrochemical potential of electrons** or **partial molar free energy of electrons** which they define in thermodynamics. We need not be concerned with the thermodynamic definitions; we only need to remember that quasi-Fermi potential introduced by Shockley in 1949 is just a neat exponential transform to help the engineers who are more familiar with electric potential and electric field and less familiar with particle concentrations and densities. The exponential transform changes the variable from the concentration of the charge carriers (electrons and holes), which the chemists know how to measure, to the electrical potential variable, which the electrical engineers know how to measure.

340 CONTINUITY EQUATION OF CHARGE AND CURRENT

One of the basic relationships that governs the generation, recombination, trapping, and flow (drift and diffusion) of electrons or holes in a solid is the equation of continuity of charge and current. It is the macroscopic **particle conservation law** that accounts for the appearance and disappearance of particles of one species in a volume element $\Delta x\Delta y\Delta z$ and time interval Δt. The **continuity equations**, one for electrons and one for holes, are given by

$$\begin{aligned}\partial n/\partial t &= g_N - r_N + q^{-1}\nabla\cdot j_N \\ &= g_N - r_N + q^{-1}[(\partial j_{Nx}/\partial x) + (\partial j_{Ny}/\partial y) + (\partial j_{Nz}/\partial z)] \qquad (340.1)\end{aligned}$$

and

$$\partial p/\partial t = g_P - r_P - q^{-1}\nabla\cdot j_P$$

$$= g_P - r_P + q^{-1}[(\partial j_{Px}/\partial x) + (\partial j_{Py}/\partial y) + (\partial j_{Pz}/\partial z)] \qquad (340.2)$$

Stating the particle conservation law in words, the change of the number of electrons in a time interval dt at time t and in a volume element of dxdydz at a point (x,y,z) in space, consists of (i) the rate of gain of electron minus the rate loss of electrons, given by (g_N-r_N)dxdydz, and (ii) the number of electrons flowing into the volume minus those flowing out per unit time, given by the divergence of the current density, $(\nabla\cdot j_N/q)$dxdydz. A straight-forward elementary derivation of these terms are given as follows.

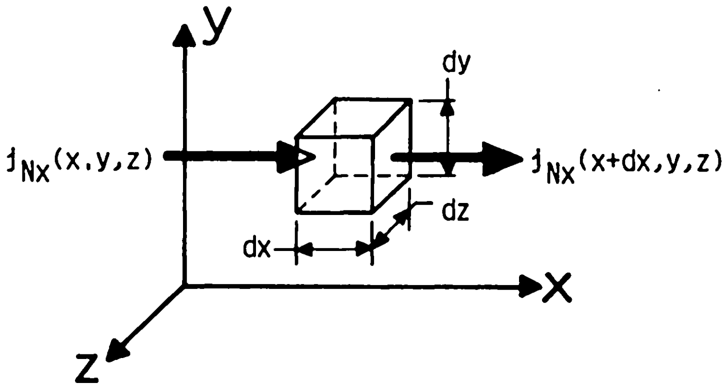


Fig.340.1 The three-dimensional coordinate system and the volume element dxdydz used to derive the continuity equation.

Let us consider the second or current divergence term first and use the volume element dxdydz shown in Fig.340.1. Denote the electron current density flowing in the positive x-direction through the y-z plane at x by $j_{Nx}(x,y,z)$, which represents electrons flowing into the volume element dxdydz; and the electron current density flowing through the y-z plane at x+dx by $j_{Nx}(x+dx,y,z)$, which represents electrons flowing out of the volume element dxdydz. The net electron current flowing out of the volume element through its two surfaces at x and x+dx with area dydz is then the difference

$$j_{Nx}(x+dx,y,z)dydz - j_{Nx}(x,y,z)dydz$$

$$= j_{Nx}(x,y,z)dydz + (\partial j_{Nx}/\partial x)dxdydz - j_{Nx}(x,y,z)dydz$$

$$= (\partial j_{Nx}/\partial x)dxdydz \qquad (340.3)$$

Since the electron carries a negative charge, the net rate of increase of the number of electrons in dxdydz due to this electron current is just the net electron current flowing out divided by the magnitude of the electron charge, q. Thus, the rate of increase of the number of electrons in this volume due to the net electron current flowing out of this volume is

$$q^{-1}(\partial j_{Nx}/\partial x)dxdydz$$

A similar derivation can be made for the electron current flowing in the y and the z directions for the general three-dimensional case.

The total rate of change of the number of electrons in this volume element also consists the electrons created by the generation and detrapping (emission) mechanisms and the electrons lost by recombination and trapping (capture) mechanisms. For example, electrons can be lost by recombination with holes directly, giving off lattice vibrations (phonons) and photons. Electrons can also be created when lattice vibrations or photons break the covalent bonds which create electron-hole pairs. Electrons can further be lost by being captured by an impurity and subsequently bound to that impurity and hence becomes unavailable for conduction. These electrons are then known as **trapped** or **captured electrons**, in contrast to the electrons in the conduction band or holes in the valence band which are sometimes called 'free electrons' although they are not truly free as in a constant potential but are continuously experiencing a periodic potential due to the host ions. A band electron can also be created by the inverse process in which an electron trapped to an impurity or defect center is released from the impurity or defect potential well by thermal vibration of the lattice (phonons) or by light (photons). This released electron is known as a **detrapped** or **emitted** electron. It increases the electron density in dxdydz.

These generation-recombination and trapping-detrapping processes will contribute to the rate of change of the number of the untrapped or 'free' electrons in the volume element, dxdydz. Let the rate of creation or generation of electrons per unit volume due to these generation-detrapping mechanisms be denoted by $g_N(x,y,z,t)$ (number/cm^3-sec) and the rate of or recombination-trapping of electrons per unit volume be denoted by $r_N(x,y,z,t)$ (number/cm^3-sec). Then the total rate of change of the number of electrons in the volume element dxdydz is given by

$$\begin{aligned}(\partial n/\partial t)dxdydz = &+ [g_N(x,y,z) - r_N(x,y,z)]dxdydz \\ &+ q^{-1}[+ (\partial j_{Nx}/\partial x)dxdydz + (\partial j_{Ny}/\partial y)dxdydz \\ &+ (\partial j_{Nz}/\partial z)dxdydz]. \end{aligned} \tag{340.4}$$

Canceling out the volume element, the continuity equation for electrons given by (340.1) is then obtained. A similar derivation can be carried out for the hole continuity equation, (340.2).

350 THE SHOCKLEY EQUATIONS FOR SEMICONDUCTORS

A summary of the six partial differential equations are given in this section. They govern the drift, diffusion, and generation-recombination-trapping-detrapping of electrons and holes in a semiconductor. For one species of traps, they are:

$$q(\partial n/\partial t) = +\ \nabla\cdot J_N + q(g_N - r_N) \tag{350.1}$$

$$q(\partial p/\partial t) = -\ \nabla\cdot J_P + q(g_P - r_P) \tag{350.2}$$

$$J_N = +\ q\mu_n nE + qD_n\nabla n \tag{350.3}$$

$$J_P = +\ q\mu_p pE - qD_p\nabla p \tag{350.4}$$

$$\nabla\cdot\varepsilon E = q(p - n + N_{DD} - N_{AA} - n_T) \tag{350.5}$$

$$(\partial n_T/\partial t) = (g_P - r_P) - (g_N - r_N) \tag{350.6}$$

The first two equations, (350.1) and (350.2), are the continuity equations of electrons and holes derived in section 340. The second two equations, (350.3) and (350.4), are the electron and hole current equations derived in (320.10) and (320.11). The fifth equation, (350.5), is the Poisson equation which relates the divergence of the electrical displacement to the macroscopic volume charge density. The sixth equation, (350.6), is the rate equation of the trapped electron concentration (or the trapped hole concentration if it is a hole trap).

We have coined these six equations the **Shockley Equations** after the late William Shockley who introduced the first four equations to analyze semiconductor devices [350.1], and who invented the p/n junction diode, the bipolar p/n junction transistor [350.2], and the junction-gate field-effect transistor [350.3] during 1948-1952 by finding the simplest analytical solutions based on simple electron and hole physics. Shockley also laid the foundation of the electron-hole recombination kinetics [350.4] which led to the formulation of the sixth equation introduced by this author in 1964 to develop the theory of low-frequency noise in junction-gate field-effect transistor [350.5]. It also led to the development and invention of the transient capacitance and current techniques by this author with his graduate students in 1967 [350.6] and the DLTS technique in 1971 [350.7], which enabled the detection and detailed characterization of minute concentrations of impurities in semiconductor device research and VLSI manufacturing.

The six Shockley equations are coupled nonlinear partial differential equations of the **three dependent macroscopic concentration variables**: the electron, hole and trapped electron concentrations, and the **four independent space-time variables**, x, y, z and t. The space-time variables are not discrete but cover a range Δx, Δy, Δz, and Δt, within which the macroscopic concentration variables are defined by averaging. The coefficients of the equations are the **macroscopic parameters: mobility, diffusivity, generation-recombination-trapping rates** and they are the value averaged over the microscopic scattering, generation-recombination, and trapping-detrapping events in the volume element $\Delta x\Delta y\Delta z$ and time interval Δt. These parameters are highly nonlinear since their values can change with position and time (x,y,z,t) in the presence of an applied force. Additional equations are necessary to give electric field dependences which are derived in advanced courses using the Boltzmann transport equation to derive the Shockley equations.

Thus far, we have described the underlying physics and the elementary derivation of the first four Shockley equations. The terms not yet discussed in detailed are the generation-recombination-trapping-detrapping terms in the first two, the fifth, and the sixth equations. The associated symbols are defined as follows: n_T for the trapped electron concentration, g_P and g_N for the generation-detrapping rate of holes and electrons, and r_P and r_N for the recombination-capture rate of holes and electrons. These terms account for the appearance and disappearance of electrons in the conduction band and holes in the valence band. There are many important fundamental energy and momentum exchange mechanisms which determine the rate of generation-recombination-trapping-detrapping of electrons and holes. These mechanisms will be systematically described in the next section, 360, and its subsections, 36np (n and p are 0,1,2,3,4,5).

Modern semiconductor device theory and numerical design methodology are based on solutions of these six coupled nonlinear partial differential equations. Analytical solutions can be obtained only for simple cases for which approximations (linearization) can be made to the dominant physical mechanisms that control the current (diffusion or drift) and the charge generation-recombination-trapping-detrapping rates. Most of these simple and zeroth order solutions were found, discovered or invented by Shockley who relied on simple physical intuition to derive the correct approximate solutions from which he invented the transistors during 1948-1952 and for which he, Bardeen, and Brattain received the Nobel Prize in Physics.

[350.1] W. Shockley, **Electrons and Holes**, Van Nostrand, 1952.

[350.2] W. Shockley, "The theory of p/n junctions in semiconductors and p/n junction transistors," Bell System Tech. Journal, 28(7), 436-489, July, 1949.

[350.3] W. Shockley, "A unipolar field effect transistor," Proc. IRE, 40(11), 1365-1376, Nov. 1952.

[350.4] W. Shockley and W. T. Read, Jr.' "Statistics of recombination of electrons and holes," Phys. Rev. 87(9), 835-842, Sept. 1952.

[350.5] C. T. Sah, "Theory of low-frequency generation noise in junction-gate field-effect transistors," Proc. IEEE, 52(7), 755-813, July 1964.

[350.6] C. T. Sah, et. al. "Thermal and optical emission and capture rates and cross sections of electrons and holes at imperfection centers in semiconductors from photo and dark junction current and capacitance experiments," Solid-State Electronics 13, 759-788, June 1970.

[350.7] Leopold D. Yau and C. T. Sah, "Thermal ionization rates and energies of electrons and holes at silver centers in silicon," phys. stat. sol. (a)6, 561-573, 16 August 1971.

360 GENERATION, RECOMBINATION, TRAPPING AND TUNNELING

The **macroscopic** generation-emission rate of electrons, g_N, and holes, g_P, and the **macroscopic** recombination-capture rate of electrons, r_N, and holes, r_P, in the continuity equations (350.1) and (350.2), and in the electron trapping equation (350.6) determine some of the most important characteristics of both bipolar and field-effect transistors. For example, the generation-emission rate determines the leakage current of silicon, compound semiconductor, Schottky barrier junction diodes, and the standby power dissipation of silicon integrated circuit chips. The recombination-capture rate can determine the speed at which a p/n junction diode or a bipolar transistor can be electrically switched because when the diode or transistor is switched into the highly conductive or on state, many minority carriers are injected into their base layer. When the diode or transistor is switched off, these injected minority carriers must be removed from the base layer before the diode or transistor is completely turned off to the low current state. These minority carriers can be removed by recombination with the majority carriers. Thus, the recombination rate will determine how fast a diode or transistor can be switched off.

When electrons recombine with holes, energy is given off. To generate electrons and holes, energy must be supplied. There are four fundamental atomic or **microscopic** mechanisms by which **recombination energy** can be dissipated and **generation energy**, supplied. These **energy exchange mechanisms** are: **(A) Thermal (phonons), (B) Optical (photons), (C) Auger-Impact (third electron or hole), and (D) Collective (plasmon).** The fourth, plasmon, mechanism involves all the 10^{23} **valence electrons collectively** supplying the generation energy or dissipating the recombination energy. These recombination and generation transitions, which change the electron (or hole) energy, are known as **inelastic**. In addition to these inelastic mechanisms, electron and hole can also recombine or be generated by **elastic tunneling** in which the electron or hole energy is constant, and the initial and final energies of the electron or hole is unchanged.

In addition to the energy exchange mechanisms, **momentum exchange** is also important in determining the transition rate or the probability of the electron-hole recombination-generation-trapping processes. The momentum exchange mechanisms can be categorized according to whether the transition involves the direct interaction of electrons with holes, or through a physically real intermediate bound state localized at a foreign impurity or intrinsic defect center. The direct-interaction transition is known as the **band-to-band** or **interband transition** since an electron in the conduction band interacts directly with a hole in the valence band when they are physically in the vicinity of each other. The two reciprocal interband transition processes are the **interband recombination of electron-hole pairs** and its inverse, the **interband generation of electron-hole pairs.**

The second group of transition processes are those involving a real intermediate step at a bound quantum state localized at an impurity or a defect center. The bound state is a real 3-d bound-electron or bound-hole solution of the Schrödinger equation in contrast to a virtual state which is associated with a second or higher order electronic transition process. The impurity or defect center that has a bound electron or hole solution is known as an **electron trap** or a **hole trap**. The model of the bound state at a trap or center is similar to that of the group-III acceptor (hole trap) and group-V donor (electron trap) discussed in chapter 2, except that the trap energy level is near the midgap rather than near the band edge. The two reciprocal electron transitions between a band state and trap (bound) state are known as the **band-to-trap** (band-to-bound) and **trap-to-band** (bound-to-band) transitions. They may be denoted as **band-trap** or **trap-band** transitions. There are four transitions for the two charge carrier types: the **electron capture** by a trap and its inverse, **electron emission** by a trap; and the **hole capture** by a trap and its inverse, **hole emission** by a trap.

The third group of transitions involves trapped electrons moving to an adjacent trap. They are denoted as **intertrap transitions**. Aside from the one-step elastic intertrap tunneling mechanism, all other intertrap mechanisms involve two or more intermediate states or steps. The intermediate state can be (i) a virtual state instead of a bound state which is a second or higher order quantum mechanical transition process, or (ii) a real bound state which consists of thermal, optical or Auger-impact transition from the initial bound state to the intermediate bound state at the center and elastic tunneling transition to the adjacent center.

To systematically analyze, study, describe and teach the transition mechanisms and kinetics of these generation-recombination-trapping-tunneling (GRTT) transitions, I have adopted a matrix table scheme in 1971 which was updated in 1986. The eighteen mechanisms are tabulated in a 3x6 (3 columns and 6 rows) table, Table 360.1. The transition mechanisms are designed by their numerical matrix element label, cr, where c=column number and r=row number.

Table 360.1
Generation-Recombination-Trapping-Tunneling Mechanisms in Semiconductors and Solids

ENERGY EXCHANGE MECHANISM		INITIAL AND FINAL STATES OF QUANTUM TRANSITIONS OF ELECTRONS AND HOLES		
		(1) Band-Band (Interband)	(2) Band-Trap (Bandtrap)	(3) Trap-Trap (Intertrap)
(0) Tunneling	(Elastic)	10	20	30
(1) Thermal	(Phonons)	11	21	31
(2) Optical	(Photons)	12	22	32
(3) Auger-Impact	(E or H)	13	23	33
(4) Tunneling	(Inelastic)	14	24	34
(5) Collective	(Plasmons)	15	25	35

The first treatment of the many transitions in one article or book was given by R. N. Hall in an article in 1968. Less than half of the fifteen mechanisms were discussed by Hall since many were not experimentally identified in semiconductors or theorized at the time. The systematic approach of using a matrix table and the IEEE symbols was first advocated and practiced by this author in 1971 [360.1] in an attempt to ease the transition for beginners and veteran alike, from semiconductor device physics to circuit representation of devices, when the equivalent circuit model was developed to analyze and compute the device response under all kinds of signal conditions (d.c., small-signal transient, small-signal sinusoidal steady-state, and large-signal analyses). It was subsequently expanded by the author in 1987 [360.2] to include the plasmon mechanism because its importance on MOS transistor reliability when operated at high voltages or exposed to ionizing nuclear radiation (x-ray, γ-ray, keV electron beam). The latest extension was given by the author in 1989 when he taught the final draft of this book to junior EE students at Florida. This final categorization consolidated all the tunneling and nontunneling mechanisms into one compact table, Table 360.1.

[360.1] C. T. Sah, "Equivalent circuit models in semiconductor transport for thermal, optical, auger-impact and tunneling recombination-generation-trapping processes," physica status solidi, (a) 7, pp.541-559, 16 October 1971.
[360.2] C. T. Sah, "Models and experiments on degradation of oxidized silicon," Solid-State Electronics, v33(2), pp.147-167, February 1990.

Some practical examples will be given in the following paragraphs for each of the generation-recombination-trapping-tunneling processes listed in the matrix table. The mathematics will be given to show how the generation and recombination rates, g_N, g_P, r_N, r_P, are related to the fundamental parameters which are the macroscopic average of the quantum mechanical rate coefficients. Also shown are their relationship to the parameters that characterize the electronic materials, such as the minority carrier recombination and generation lifetimes which have so often been introduced only empirically to characterize the switching speed and leakage current of diodes and transistors.

A judicial choice of notations or symbols is crucial to facilitate presentation and description of known GRTT mechanisms and to discover new mechanisms. It is also important for the successful communication of results verbally and in writing (textbooks, reference books, research articles). Proper choice of symbol, subscript and superscript is mandatory to make the physics represented by the symbol transparent to the readers and students. The symbol must be easily recognized and understood at one glance, and not add confusion to the many facets of the physical phenomena, of the nonlinear functional dependences, and of the tedious circuit representations of the characteristics of semiconductor devices. The symbol must also be selected to help the users to recall the parameter's physics and engineering meaning so that the users' memory is reserved for reasoning, analyzing and inventing. We have already tacitly followed these criteria by adhering to the IEEE Standards of symbols for circuits and devices in the previous chapters. The IEEE Standards of symbols is the choice since it was developed by a committee of veteran researchers, teachers and practicing engineers with deep, far-reaching insights and broad anticipations to cover all possible current engineering practices and future state-of-the-art developments.

In applying the IEEE symbols to the generation-recombination-trapping mechanisms, we use the symbol g for generation rate and r for recombination rate of electron-hole pairs; c for the capture rate of an electron or a hole by a trap (an impurity or defect center which can bind an electron or a hole); and e for emission (excitation, ejection or release) of the bound electron or hole from the trap. Subscript n denotes electron and subscript p denotes hole. Superscript denotes the physical mechanism that controls the rate of the generation, recombination, capture and emission transition processes. For example, the instantaneous interband thermal generation rate of electrons, designated as process 11 in Table 360.1, would be denoted by $g_N{}^t(r,t)$ where the instantaneous value is indicated by a lower case symbol, g, and an upper case subscript, N. Lower case subscript n will be used to denote a rate coefficient which is assumed constant and usually stays approximately constant in a given problem if deviation from thermal equilibrium is small or the applied electric field is not high so hot electron effects are unimportant. The particle involved in the transition, electron, is denoted by the subscript N while the energy exchange mechanism involved, thermal, is denoted by the superscript t. We have an additional option of an upper and lower superscript which will be will be left open for a future extension in studying the advanced theory of devices and materials. The space dependence of the generation rate is denoted by the position vector, r, while the time dependence is denoted by t in the two variables enclosed by the parenthesis following a symbol, (r,t). Similarly, $g_P{}^o(r,t)$ is the instantaneous interband optical generation rate of holes designated as process 12 and $g_P{}^n(r,t)$ is the instantaneous interband impact generation rate of electrons by hole impact designated as process 13. The kinetics of selected processes in Tables 360.1 will be discussed in the following subsections numbered by 36cr. c is the column number which denotes the initial and final states of the electronic transition and the momentum exchange mechanism, and r is the row number which denotes the energy exchange mechanism of the electronic transition.

3611. Interband Thermal Generation and Recombination

In this interband (band-to-band) thermal process labeled 11 in Table 360.1, electrons and holes recombine in pairs and are generated in pairs so that

$$g_{11} = g_N{}^t = g_P{}^t \qquad (3611.1)$$

and

$$r_{11} = r_N{}^t = r_P{}^t \qquad (3611.2)$$

These transitions are illustrated in the energy band diagrams shown in Figs.3611.1(a) and (b). They are known as the transition **energy-band diagrams.** They show the **initial** state of each participating particle just prior to the transition. The **arrows** indicate the **direction** of **electron transition**. They are opposite to the direction of hole transition. In order to avoid confusion, the arrow direction will be used to denote the direction of electron transition exclusively in this mechanism. A second transition energy-band diagram is needed to show the final states after the transition so as not to be confused with the initial states.

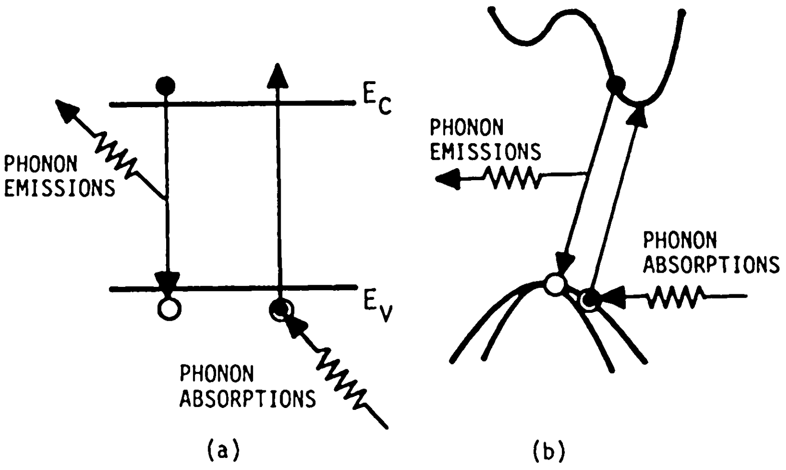


Fig.3611.1 The initial state transition energy band diagrams of the interband thermal recombination and generation transitions. (a) The E-x and (b) E-k diagrams.

In an interband thermal recombination event, the energy lost by the electron while dropping into a hole is carried away by setting off the surrounding silicon ions to vibrate about their equilibrium positions. The outward propagating lattice vibration waves (or phonons) then dissipate the **recombination energy** as heat which is passed on to the surrounding ambient by the vibrating host atoms on the solid's surface. In the inverse process, i.e. the electron-hole pair generation process, the lattice vibrations or phonons act in concert to break or rupture a covalent bond and to shake loose, release or free an electron from the covalent bond. The released electron is moved to the space between the bonds and atoms, leaving a hole behind in the ruptured covalent bond as illustrated by the valence bond diagram given in Fig.171.2(a).

These two interband transitions are intrinsic generation-recombination processes of the semiconductor. Although they are not the dominant processes (12, 13, 21 and 22 are), they determine the thermal equilibrium concentration of electrons and holes, n_i, in an intrinsic or pure crystalline semiconductor. The transition rates are very small because the largest phonon energy in a solid (the optical phonons) is only about 60 meV (milli-electron-volt) while the energy gap of silicon is about 1.2 eV or 1200 meV. Thus, each transition would require as many as 20 phonons simultaneously whose probability of occurrence is small, giving small transition rates and low values of n_i.

3612. Interband Optical Generation and Recombination

In these band-to-band processes, labeled 12 in Table 360.1, a photon supplies the necessary energy to break the covalent bond, releasing an electron from the covalent bond and leaving a hole behind. An **electron-hole pair** is generated, or an electron and a hole are generated simultaneously. In the inverse process, an electron recombines with a hole and a photon is generated to carry away the recombination energy. Since the transition involve an electron and a hole simultaneously, the electron generation rate is equal to the hole generation rate,

$$g_{12} = g_N{}^o = g_P{}^o \tag{3612.1}$$

Similarly, pair recombination means that the recombination rate of an electron with a hole is equal to the recombination rate of a hole with an electron,

$$r_{12} = r_{12N}{}^o = r_{12P}{}^o. \tag{3612.2}$$

The interband optical generation of electron-hole pairs is also known as interband or intrinsic photoexcitation of electrons, and **interband optical absorption** or **intrinsic absorption**. The interband optical recombination of electron-hole pairs is also known as interband photon emission or intrinsic photon emission, and **interband radiative recombination** or **intrinsic radiative recombination**. Each of these terms emphasize one or more physical mechanisms. These optical transitions are the principal mechanisms that are responsible for the operation of many opto-electronic or photonic devices. Intrinsic absorption is responsible for the operation of: photoconductors made of crystalline semiconductors, photocapacitors, solar cells, photovoltaic diodes, and phototransistors. Radiative recombination is responsible for the operation of light emitting diodes (LED) and p/n junction laser diodes.

Some materials have high optical efficiency or high **quantum efficiency** (GaAs), known as **direct** energy gap material, while others do not (Si, Ge), known as **indirect** energy gap material. The reasons are illustrated by the energy band diagrams in Figs.3612.1(a) and (b). Figure 3612.1(a) is the E-k diagram of an indirect energy gap semiconductor such as silicon and germanium. Note that both the radiative recombination and optical generation transitions involve a large change of the wave number k or the particle momentum. The change is of the order of $2\pi/a \approx 6/a$ since the E-k diagram spans $k = -\pi/a$ to $+\pi/a$. 'a' is the lattice constant which is of the order of 5×10^{-8} cm or 5 Angstrom. We do not have a problem on energy conservation since the electron energy change is exactly equal to the energy of the photon. But, we do have a problem of momentum conservation or conservation of the wave number k (momentum is equal to $\hbar k$ from de Broglie's hypothesis) during a radiative recombination event of an electron with a hole in Si.

The reason is that photons carry negligible momentum compared with the electron momentum change during the transition in a crystal. Stated differently and equivalently, the photon wavelength λ, which gives the photon momentum $\hbar k_{photon} = 2\pi\hbar/\lambda$, is much larger than the lattice or inter-atomic spacing, a, i.e., $\lambda >> a$. For Si, $\lambda(\hbar\nu = E_G = 1.2eV) = 1\mu m = 10^4 A$ while $a = 5.43A$. Thus, we have $\lambda = 10^4 A >> a = 5.43A$.

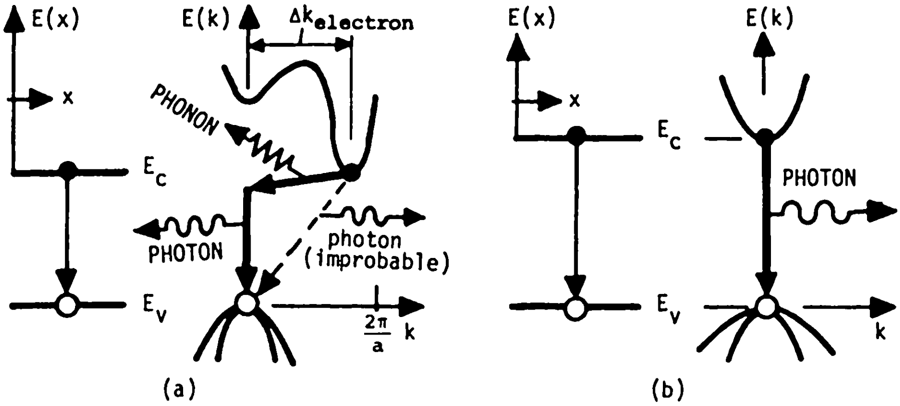


Fig.3612.1 The initial state E-x and E-k energy band diagrams of interband radiative recombination transitions showing the states of the electrons and holes before the transitions and the arrows indicate the direction of electron transition. (a) Interband radiative recombination in an indirect semiconductor such as Si ($E_G = 1.2eV$). (b) Interband radiative recombination in a direct semiconductor such as GaAs ($E_G = 1.45eV$).

A numerical calculation showing $\lambda = 10^4 A$ will now be described. The photon wavelength can be computed from the photon energy using Planck's hypothesis. The photon energy which must be equal to or somewhat greater than semiconductor energy gap since electrons and holes in the semiconductor are near the conduction and valence band edges, E_C and E_V, respectively. Using the Planck's relationship between energy and frequency of oscillation, $E = \hbar\omega = h\nu = hc/\lambda$, where we have used $\nu\lambda = c$ where ν, λ and c are respectively the frequency, wavelength, and velocity of light. Inserting numerical values of h and c, the wavelength of a photon λ (in μm) as a function of photon energy E (in eV) is then given by

$$\lambda = hc/E = 1.24eV/E \quad \mu m \tag{3612.3}$$

$$= 12400eV/E \quad A. \tag{3612.4}$$

Thus, the energy of the photon, given off during a radiative electron-hole recombination event in silicon (energy gap is 1.2 eV), is equal to or slightly more than 1.2 eV. The wavelength would be about 10,000 Angstroms or 1 micrometer

(1 μm). Recall that the human eye is most sensitive to yellow light whose wavelength is about 5000A and whose photon energy is about 2.5 eV. Thus, the weak light emitted from radiative electron-hole recombination in Si (1.2 eV and 10,000A) is in the far infrared range, invisible to human eyes.

Furthermore, the number of electrons and holes in silicon undergoing radiative or light-emitting recombination is very small owing to silicon's **indirect** energy band structure. The indirect energy band causes a large change in electron momentum or wave number (k) of the recombining electrons and holes. This large change cannot be carried away or supplied by photons. To illustrate this **momentum conservation bottleneck**, a simple numerical calculation is now given.

The momentum carried away by the photon can be obtained using de Broglie's relationship

$$P_{photon} = h/\lambda = 4.1357\text{x}10^{-15}(\text{eV-sec})/10^{-4}(\text{cm})$$
$$= 4\text{x}10^{-11}\ \text{eV}/(\text{cm/sec}).$$

However, the electron moving into or recombining with a hole directly would change its wave number by $\Delta k \approx 2\pi/a$ as evident from the Si E-k diagram given in Fig.3612.1(a). For Si, this would give a electron momentum change of

$$\Delta P_{electron} = \hbar\Delta k \approx \hbar(2\pi/a) = h/a \qquad (3612.5)$$
$$= 4.1357\text{x}10^{-15}(\text{eV-sec})/10^{-8}(\text{cm}) \qquad (3612.5\text{A})$$
$$= 4\text{x}10^{-7}\ \ \text{eV}/(\text{cm/sec}) \qquad (3612.5\text{B})$$

Thus, $\Delta P_{electron} = 10^4 P_{photon} >> P_{photon}$, i.e., the change of the electron momentum during an interband electron- hole recombination event is 10,000 larger than momentum that can be carried away by the photon because the lattice constant, a, is 10^4 times larger than wavelength of the energy-gap photon.

In order for an interband radiative recombination event to proceed in Si and other indirect materials, there must be another particle to carry away the large change of electron momentum. This particle is usually a phonon whose momentum or wave number spans the same range in k-space as that of the electron, about $2\pi/a$. This would make the recombination process a three-body rather than a two-body initial-state transition (3-body: electron, hole and phonon; 2-body: electron and hole; or 4-body and 3-body final states counting the photon). The probability of a 3-body transition is much smaller than a 2-body transition. Hence, in an indirect semiconductor such as Si, the intrinsic radiative recombination (interband optical recombination) and its inverse intrinsic absorption (interband optical generation) have also a rather small probability and low rate. This is reflected by the low values of $g_N{}^o$, $r_N{}^o$, $g_P{}^o$ and $r_P{}^o$ in Si.

The interband optical transition rates are significantly greater in materials with a direct energy-wave number diagram, such as GaAs, illustrated in Fig.3612.1(b). For this material, the conduction band minimum or conduction band edge is positioned at the same k value as the valence band maximum or valence band edge. Both are at k=0 in this figure. Thus, when an electron in the conduction band drops into a hole in the valence band, there is very little change of the electron momentum or k value. Consequently, both energy conservation and momentum or wave number conservation are satisfied by the emission of a photon. There is no need for a third particle, such as a phonon, to carry away the momentum change since there is no change in electron and hole momenta. Thus, the optical transition processes in a direct energy gap material are very effective and have high transition rate. This is the fundamental reason that GaAs gives efficient LED (Light Emitting Diode) and p/n junction diode injection lasers.

The low radiative recombination rate in indirect material, of the order of one per millisecond in Si and Ge, means that the minority carriers can have relatively high recombination lifetimes. Hence, very high current gains can be obtained in bipolar transistors which depends on the minority carriers to carry the to-be-amplified signals. The high radiative recombination rate in direct materials, of the order of one nanosecond in GaAs, means that the minority carrier lifetime is very short, making it difficult to produce high gain bipolar transistor in GaAs. However, because of this fundamental difference, very efficient light emitting diodes can be made from GaAs while practically no light is generated by a Si diode.

3613. Interband Auger Recombination and Impact Generation

These band-to-band transitions, labeled 13 in Table 360.1, utilize a third material particle (an electron or a hole) to carry away the recombination energy in an electron-hole recombination event. This three-particle initial-state transition is known as **interband Auger recombination.** In the electron-hole generation transition, an energetic electron or hole is present to knock an electron out of the covalent bond by impact, giving three particles. This three-particle final-state transition is known as **interband impact generation.** The generation rates of electrons and holes by electron impact are equal since the electrons and holes are generated in pairs. Thus,

$$g_{13}{}^{n} = g_{N}{}^{n} = g_{P}{}^{n}. \qquad (3613.1)$$

Similarly, the interband impact generation rates of electrons and holes by hole impact are also equal and given by

$$g_{13}{}^{p} = g_{N}{}^{p} = g_{P}{}^{p}. \qquad (3613.2)$$

The Auger recombination rates have similar relationships, given by

$$r_{13}{}^{n} = r_{N}{}^{n} = r_{P}{}^{n} \qquad (3613.3)$$

and

$$r_{31}{}^{p} = r_{N}{}^{p} = r_{P}{}^{p}. \qquad (3613.4)$$

Figures 3613(a)-(d) give E-x energy band diagrams showing the initial electron and hole states of the interband Auger-impact transitions. Corresponding E-k diagrams can be readily drawn and are left as an exercise.

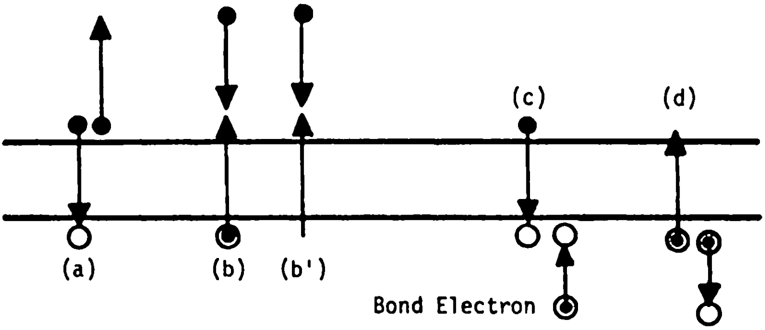


Fig.3613.1 Interband three-body transitions. (a) and (c) Interband Auger recombination of electron with hole. (b), (b') and (d) Interband impact generation of electron-hole pair.

Figures 3613.1(a) and (c) show that the interband Auger recombination process requires the presence of two electrons and one hole, or two holes and one electron, in the vicinity of each other. Their inverse are shown in figures (b) and (d). A new notation, the circled dot, is introduced to represent the valence bond electron that is undergoing a transition. Its necessity is indicated by the non-descript lower arrow in figure (b') or a similar one that would have appeared in (c). This is further extended later by using a hollow-head arrow for hole transitions and solid-head arrow for electron transitions.

The three interaction forces in figure (a) are the two attractive Coulomb forces between the hole and the two electrons and the one repulsive Coulomb force between the two electrons. Energy conservation is maintained via these three forces as one electron drops into a hole and the recombination energy ($\geq E_G$) is carried away by the second electron via Coulomb repulsion. The probability of having two electrons next to each other is high if the electron concentration is high. High electron concentration can come from a high concentration of donor dopant impurity. Similarly, a high concentration of holes is necessary to have a high probability of two holes next to each other as indicated by Fig.3613.1(c). Thus, the interband Auger recombination processes become important in highly doped lo-resistivity regions such as the emitter layer of a diffused bipolar transistor or a p/n junction diode, or in a layer exposed to intense intrinsic light, ($h\nu > E_G$), x-ray, or energetic particles (keV electrons or nuclear radiation) which can generate a very high concentration of electrons and holes. Large interband Auger recombination rate lowers the minority carrier lifetime in the layer. For example, the current gain of bipolar transistors is reduced if their emitter layer has a very high concentration of dopant impurity.

The inverse three-particle interband Auger recombination transition is the interband impact generation of an electron-hole pair by an energetic electron or hole indicated in Figs.3613.1(b) and (d). These mechanisms cause the current in p/n and metal/semiconductor diodes to become infinite at a large and unique applied reverse voltage in the absence of a current limiting resistance.

The initial electron or hole energy required to break a covalent bond in order to create an electron-hole pair can be calculated using a very simple model formulated by Shockley in 1961 known as the **ballistic model** (particle) or **spike model** (spike in spatial distribution of electron number of density). Consider the transition diagram given in Fig.3613.1(b) where an energetic electron knocks out a valence bond electron to create an electron-hole pair. Obviously the initial electron must have a kinetic energy at least equal to the energy gap because of energy conservation. However, because of the momentum or k conservation requirement, the initial kinetic energy required to break a covalent bond is significantly larger than the energy gap. If the electron and hole masses are equal, then the initial electron must have a kinetic energy 50% greater than the energy gap, $(3/2)E_g$, to generate an electron-hole pair in order to also conserve momentum or k. This is the **threshold energy of electron** (and **threshold energy of hole**) for interband impact generation of electron-hole pairs. They are known as the **electron and hole impact ionization thresholds**. The term 'ionization' was used by the first researchers who investigated the p/n junction breakdown as an analogy of the same phenomenon in a gas-filled tube which causes the gas to ionize, such as in the neon light bulb. For equal electron and hole effective masses, this ionization threshold energy is $(3/2)E_G=(3/2)\mathrm{x}1.2=1.8$ eV for Si, and $(3/2)\mathrm{x}9=13.5$ eV for SiO_2.

The factor 3/2 can be simply derived using classical mechanisms applied to a three body collision containing the initializing electron, and the created electron and hole. The simplification comes from the equal effective mass assumption. The minimum or threshold energy to satisfy both energy and momentum conservation is the situation in which the initial kinetic energy and the initial momentum carried by the impacting energetic electron are equally shared by the two electrons and a hole after the collision. Denote the initial kinetic energy as E_I, the initial momentum as P_I where $E_I=P_I^2/2m$, and the final kinetic energies and momenta of the three particles as E_1, E_2, E_3, P_1, P_2, and P_3, then

$$E_1 = E_2 = E_3 \tag{3613.5}$$

and

$$P_1 = P_2 = P_3 = \surd(2mE_1) = \surd(2mE_2) = \surd(2mE_3). \tag{3613.6}$$

The energy and momentum conservation equations are

$$E_I = E_1 + E_2 + E_3 + E_G \tag{3613.7}$$

and

$$P_I = P_1 + P_2 + P_3. \tag{3613.8}$$

For the present case of equal effective mass case, these reduce to

$$E_I = 3E_1 + E_G \tag{3613.9}$$

and

$$\surd E_I = 3\surd E_1. \tag{3613.10}$$

The solution by inspection is

$$E_1 = E_2 = E_3 = E_G/6 \tag{3613.11}$$

and

$$P_1 = P_2 = P_3 = \sqrt{(mE_G/3)} \tag{3613.12}$$

which give

$$E_I = (3/2)E_G \tag{3613.13}$$

and

$$P_I = \sqrt{3mE_G}. \tag{3613.14}$$

For three-dimensional energy bands such as Si, there are directions of low threshold energy due to the directional dependences of the effective masses and the indirect energy band structure. Both affect momentum conservation. The threshold in these low threshold directions could be reduced from $3E_G/2$ to a value just slightly larger than the energy gap. These directions are the preferred channels along which the electron-hole pairs will pass through during the interband impact generation events.

Interband impact generation process is important in a reverse-biased p/n junction diode. The reason is the high electric field and large potential drop across the junction space charge layer when a high voltage is applied to the diode terminals in the reverse current flowing direction. The large electric field accelerates the electrons and holes to a kinetic energy greater than the impact threshold energy to initiate the electron-hole pair generation process. The applied reverse voltage will all appear across a very thin space-charge layer in the p/n junction. The thinness of this high electric field layer makes it possible to accelerate the electrons to a kinetic energy greater than the impact generation threshold, 1.8 eV, before the electrons have a chance to lose their kinetic energy by inelastic phonon scattering.

Interband impact generation of electron-hole pair in a high electric field is one of the many **hot electron, hot hole, or hot carrier effects**. One of them, the drift velocity saturation effect, was discussed in section 314. The effective electron temperature during the interband impact generation process can be computed as follows. At room temperature (300K or 22C), the average electron kinetic energy is about $(3/2)kT = 1.5 \times 0.026 = 0.039$ eV. The kinetic energy of the electron that initiates an interband impact generation of electron-hole pair must be greater than the threshold energy which is about $(3/2)E_G = 1.5 \times 1.2 = 1.8$ eV in Si. Thus, this 1.8eV electron would have an equivalent electron temperature of $T_{electron} = 300K \cdot (1.8eV/0.039eV) = 300 \cdot 46.15 = 14{,}000K$ or about 50 times hotter than a room-temperature or thermal electron.

3621. Band-Trap Thermal (SRH) Generation-Recombination-Trapping

This is labeled 21 in Table 360.1. It is known as the **Shockley-Read-Hall (SRH)** generation-recombination-trapping (GRT) process. When impurities from foreign chemical elements or lattice defects (such as a missing host atom, an interstitial or misplaced host atom) are present in a semiconductor, a large change of the crystal's periodic potential may be created which is localized or centered around the impurity or defect. If the localized potential change is large enough, it can bind or trap an electron or a hole. The electron and hole traps are known as centers following the traditional usage of 'color centers' in ionic or alkaline crystals or salts. The trapping properties of the group III and V substitution impurities in semiconductors were briefly described in chapter 2. The atomic hydrogen model and the dielectric screening concept were used to demonstrate the magnitude of the binding energy and the radius of the ground-state Bohr orbit of the trapped electron or hole. Similar trapping properties can be expected for the impurity elements from columns I, II, VI and even VII of the periodic table. Their nuclear composition and core charge distribution are more different from that of the column IV host atom than the column III and V impurities, resulting in deeper, larger, and more localizing potential well to trap an electron or a hole. Thus, bound states with larger binding energy for electron or hole are expected from the column I, II, VI and VII impurities; their energy levels are deeper in the energy gap, usually around the midgap energy of semiconductors or deeper. For example, experiments have shown that two energy levels are present in the energy gap of silicon when gold impurity is introduced into silicon. The electrons and holes can be trapped or recombine at these two Au levels in Si. Electrons and holes can also be generated at these two Au levels. The recombination and capture energies are carried away by phonons or the vibrating Si ions surrounding the Au impurity.

Transitions at the gold center just described is an example of a recombination process localized at a recombination center or trap. It is in contrast with the interband recombination processes (11, 12 or 13) which can occur at any location in the crystal and hence are not localized or are extended or nonlocalized over the entire crystal. The term 'delocalized' used by some theorists has a different physics. The inverse process in which an electron and a hole is generated at an impurity or defect center is also a localized process in contrast with the nonlocalized or extended interband generation processes, 11, 12 and 13.

Recombination of electrons and holes at an impurity or defect center is the most important mechanism that influences the lifetime of electrons and holes. The switching speed of a p/n junction diode is improved by increasing the concentration of a specific recombination impurity. The current gain of a bipolar transistor can be improved by reducing the concentration of another specific recombination impurity. Similarly, the generation of electrons and holes at an impurity or defect center is the dominant mechanism that controls (increases) the leakage current in junction diodes and transistors, and the required refresh-rate of dynamic random access memory (DRAM) cells. Only a minute amount of impurity, such as one part per billion of gold in silicon, will greatly increase the electron and hole generation rates and thus, drastically increase the leakage current and the required DRAM refresh frequency. Numerical examples will be given in chapters 5 and 6.

The first important characteristics of the GRT center noted by device technologists was the location of electron energy level: the highest leakage current in Si p/n junction diodes were usually from GRT levels near the middle of the energy gap. This larger depth of the energy level, caused by the deep potential well of the GRT centers, gave the popular name **deep levels**. In contrast, the energy level of the group III (B,Al,Ga) and group V (P,As,Sb) dopant impurities are near a band edge. Thus, they are known as **shallow levels**. The shallow levels are generally not very effective generation-recombination centers compared with the deep levels. This is because the energy released during the recombination of a hole with a trapped electron at a shallow donor level is nearly equal to the energy gap, E_G, while the recombination energy is only about half the energy gap, $E_G/2$, when recombination occurs at a deep or midgap level. The larger recombination energy at the shallow level is harder to dissipate since it requires the emission of many more phonons than the smaller recombination energy at the deep or midgap levels. Thus, deep (midgap) levels are more efficient generation-recombination centers than shallow levels.

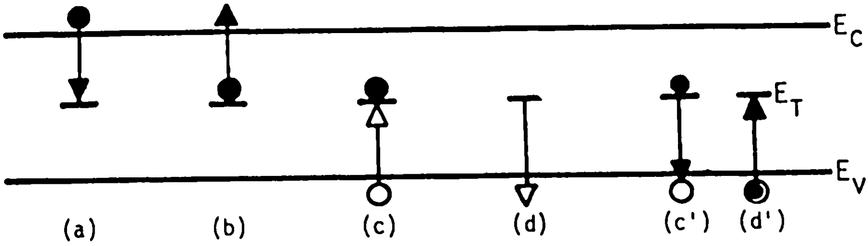


Fig.3621.1 The four transition processes at an impurity or defect center (deep level): (a) electron capture, (b) electron emission, (c) hole capture and (d) hole emission; or (c') electron emission into the valence hole and (d') capture of a valence electron. Energy conservation can be effected either by phonon or photon emission or absorption.

The four thermal electron-hole transition processes at impurity and defect centers can also be illustrated using the energy band diagram. Only the E-x diagram will be used which is given in Fig.3621.1 for an electron trap, i.e. a center that has an electron bound state. A similar diagram can be constructed for a hole trap. Process (a) is an electron capture transition in which an electron from the conduction band is captured by an empty center or unoccupied electron trap. Process (b) is just the inverse of (a) in which a trapped electron is released or emitted to the conduction band. In both transitions (a) and (b), energy conservation is effected by thermal vibrations of the host atoms or phonons. Phonons are emitted during electron capture to carry away the recombination energy. Phonons are absorbed to provide the necessary energy to emit or release the trapped electron. Figures (c) and (d) are the corresponding hole capture and hole emission transitions, depicted by the arrows showing the direction of hole transition. Process (c) is the recombination of a hole with the trapped electron which can also be viewed at the transition of a trapped electron down to a unoccupied electron state in the valence band as depicted by Figure (c'). Process (d) is the inverse of (c) but the diagram is confusing and the arrow, non-descript. A better picture is Figure (d') which shows a valence electron being excited into the trap level, E_T. This valence electron in the process of making a transition is represented by the new symbol, a circled dot.

Since the final electron state during capture and the initial electron state during emission at a trap are bound states and not free or Bloch electron state in a periodic potential, the k-conservation or momentum-conservation selection rule does not hold on the electronic transitions at a trap. However, the transition probability does depend on whether the electron bound state at the trap is more closely associated with the conduction band (and band electrons) or the valence band (and band holes). The first-principle quantum theory to predict the energy level positions is yet to be developed because it is difficult to model the screening of the large impurity or defect potential by the many core and valence electrons. It is even more difficult to predict the capture and emission rates because the trapped electron interacts with many phonons simultaneously making the problem highly nonlinear and not accountable by the linear phonon model. However, some accurate experimental data on the thermal transition rates and fairly complete experimental data on the energy level positions have been obtained for impurity and defect centers in Si, Ge, GaAs, and some $Al_xGa_{1-x}As$.

The kinetics of the capture and emission processes of electrons and holes at a trap were first discussed and modeled comprehensively by William Shockley and W. T. Read, Jr. in their classic 1952 Physical Review article [3621.1] and by R. N. Hall in a short presentation at a 1952 American Physical Society meeting [3621.2]. They are known as the **Shockley-Read-Hall** or Hall-Shockley-Read recombination statistics with the acronym **SRH recombination statistics**. The word 'statistics' employed by Shockley and Read is a misnomer because they presented a detailed derivation of the kinetics of recombination and generation of electrons and holes under an arbitrary steady-state condition. Thus, it is more appropriate to call the **SRH recombination-generation-trapping** theory, a **kinetics** or **SRH kinetics**, which Hall had done. The SRH kinetics are in fact the electronic analog of the atomic and molecular reactions in chemistry and biochemistry. They share many similar concepts and mathematical procedures as well as analytical solutions.

Associated with the SRH kinetics at a trap is the frequent usage of the terms **generation**, **recombination**, and **trapping** without an explicit definition of what these transitions or events are. The concise quantitative definitions were given by Sah and Shockley [3621.3] in terms of the macroscopic transition mechanisms shown by the microscopic transition energy band diagram given in Fig.3621.1. The generation, recombination, and trapping are events each consisting of a sequence of two consecutive capture and/or emission transitions. The four possible events on one grt center are described as follows.

(i) An electron-hole **recombination** event occurs when electron capture [transition (a)] is followed by hole capture [transition (c)].

(ii) An electron-hole **generation** event occurs when electron emission [transition (b)] is followed by hole emission [transition (d)].

(iii) An **electron trapping** event occurs when electron capture [transition (a)] is followed by electron emission [transition (b)].

(iv) A **hole trapping** event occurs when hole capture [transition (c)] is followed by hole emission [transition (d)].

Thus, an impurity or defect center with a bound state in the semiconductor energy gap can be (i) an **electron-hole recombination center**, (ii) an **electron-hole generation center**, (iii) an **electron trap**, or (iv) a **hole trap**, depending on which of the four sequences given above has the highest probability of occurrence. For example, one impurity species in a p/n junction diode may possess all four properties, but only one property dominates at one location in the diode and at one applied voltage. The quantitative criteria for a center to assume each of the four properties were delineated by Sah and Shockley in 1958 [3621.3]. They were necessary to systematically analyze the electron-hole GRT kinetics in semiconductors containing many species of impurity and defect centers, each having many energy levels and many charge states, now termed as the many-electron, many-hole, and many-electron/many-hole traps in this book.

[3621.1] W.Shockley and W.T.Read, "Statistics of recombination of holes and electrons," Phys.Rev. 87(9), 835-842, Sept.1952.
[3621.2] R.N.Hall,"Germanium rectifier characteristics," Phys.Rev.83, 228, 1951.(abstract).
[3621.3] C.T.Sah and W.Shockley, "Electron-hole recombination statistics in semiconductors through flaws with many charge conditions," Phys.Rev.109(4),1103-1115,Feb.15,1958.

3622. Band-Trap Optical Generation-Recombination-Trapping

The SRH recombination kinetics just described is also applicable to the band-trap optical transitions in which a photon provides energy conservation. These are labeled by 22 in Table 360.1. Capture of an electron (or hole) by a trap with the emission of light is known as the **optical capture** and also **radiative capture**. It is a particularly important process in applications. In light emitting diodes made of indirect gap semiconductor (GaP and others), light emitting efficient is still significant because of radiative recombination through a shallow impurity level.

The inverse, optical emission of a trapped electron (or hole), is known as **extrinsic absorption**. Semiconductor infrared detectors employ the extrinsic absorption by trapped electrons and sometimes holes to get their detection sensitivity. There are many ways to use the optical transitions at shallow and deep levels in solids to process light signals, generally called photonic or optoelectronic devices.

3623. Band-Trap Auger Capture and Impact Emission

These transitions, labeled by 23 in Table 360.1, involve an electron or a hole to supply the emission energy or to dissipate the capture energy. There are a total of 8 possible electron and hole transition processes shown in Fig.3623.1. For example, two electrons are near an electron trap. One of the two electrons is captured by the trap and the capture energy is impacted to the second electron increasing the kinetic energy of the second electron. For a second example, an electron and a hole are near an electron trap. The electron is captured by the trap and the capture energy is imparted to the hole, increasing the kinetic energy of the hole. These are the transitions of **Auger capture of an electron** by a trap. The inverse processes are known as **impact release (or impact emission) of a trapped electron** by an energetic electron or an energetic hole. These processes could become very important in practice. For example, the high energy electrons required to knock out a trapped electron at the gold level can be provided by reverse biasing a gold-doped or gold-diffused Si p/n junction diode to a high reverse voltage. Auger capture capture of a minority carrier by a trap has a linear kinetics and hence can be important in heavily doped emitter layer of a bipolar transistor.

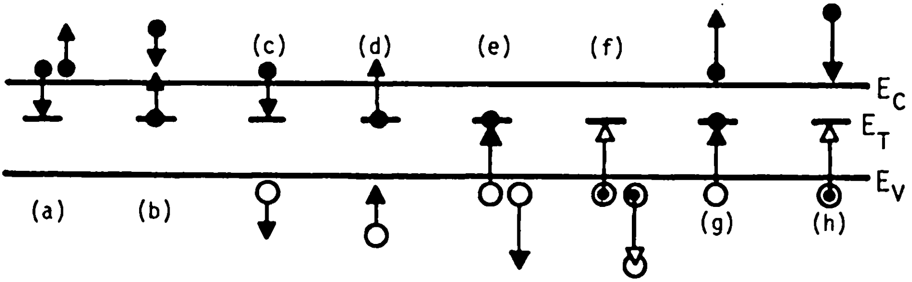


Fig.3623.1 Initial-state transition energy band diagram of the eight band-trap Auger-impact capture-emission processes.

363n. Three Intertrap Transitions

These three intertrap transitions, labeled 31, 32, and 33 in Table 360.1, do not occur frequently in semiconductor devices. The radiative intertrap transition, 32, is an exception and whose light emission has been known as the **pair spectra** since it involves the recombination of a electron trapped at a donor with a hole trapped at a near-neighbor acceptor. The donor-acceptor pair spectrum has been observed in ZnS for many (~30) years but their precise transition mechanism was identified only recently (~20 years ago) by P. J. Dean in the photon emission spectra of GaP which contained high concentrations of shallow donors and acceptors. This identification of the mechanism greatly enhanced our understanding of the energy exchange mechanisms and electron transitions in light emitting diodes (LED) and p/n junction lasers. Later on, it was also observed in Si intentionally doped with Li donor and In acceptor.

In these 32 transitions, an electron trapped to a shallow donor energy level near the conduction band edge **recombines** with a hole trapped to a nearby shallow acceptor energy level near the valence band edge. The recombination energy of the trapped electron with the trapped hole is emitted as light. The photon energy shows many peaks with many series of systematic energy separations owing to the regularity of the inter-acceptor-donor potential wall heights. This regularity arises from discrete lattice sites or discrete separation distances between the nearest neighbor, second nearest neighbor, and third and distant neighbors of the substitutional acceptor and donor centers. The spectra was coined the **pair spectra** since the radiative recombination occurs at acceptor-donor pairs.

36n0. Elastic Tunneling

The band-band, band-bound, and bound-bound elastic tunneling transitions are labeled 10, 20 and 30 in Table 360.1. The numeral 0 is used to indicate that there is zero electron energy change between the initial and final states. The three elastic tunneling mechanisms are shown in the initial-state energy band diagrams given in Figs.36n0(a), (b) and (c) where the potential barrier can come from a large-gap insulator (SiO_2) between two small-gap semiconductors or the space-charge layer of a p/n junction.

Tunnel current was first identified around 1950 by K.G. McKay and coworkers at Bell Labs. in p/n junction diodes on very low resistivity p-type and n-type Si whose reverse voltage at current-breakdown (current goes to infinity) is less than about 10V. Negative resistance due to conduction-valence interband tunneling [transition LE depicted in Fig.36n0(a)] in the forward bias direction of Ge p/n junction was identified in 1957 by Leo Easki but it was not a successful commercial product even though the negative resistance persists down to 1ps. The two-step conduction-band/trap/valence-band tunneling via an impurity level (Au in Si) such as those labeled 'excess' in Fig.36n0(b) was shown in 1958 by this author as the mechanism that gave excess current and lowered negative conductance in Esaki tunnel diodes. Tunnel transition between semiconductor band states and surface traps [transition AM depicted in Fig.36n0(b) and other interface states in the Si energy gap] was proposed mechanism of 1/f noise by Alan L. McWhorter in his doctoral thesis at MIT in 1949 which was experimentally verified by this author and his graduate students [See Physical Review Letters,17,956-958,1966; and IEEE Transaction on Electron Devices,ED-19,273-285,1972]. Recently, there has been a renewed interest on surface 1/f noise in wide-band low-noise applications of Si analog MOS integrated circuits. Since 1980, tunneling transitions (both interband and Si-band/oxide-trap) have become important aging and failure mechanisms of submicron Si MOS transistors and integrated circuits. [See a review given by this author in Solid-State Electronics, 33(2),147-167,1990.] A historical first measurement of the trap density-of-state in SiO_2 was reported by Scott E. Thompson and Toshikazu Nishida, "Tunneling and thermal emission of electrons from a distribution of shallow traps in SiO_2," Applied Physics Letters, 58, 1262-1264, 25 March 1991 using tunnel transitions out of oxide traps labeled by PS/TN in Fig.36n0(b). Furthermore, inter conduction band electron tunneling transition, Fig.36n0(a), is the principal mechanism of operation of the commercially successful EEPROM chip (electrical erasable programmable read only memory), known as flash EPROM because it can be quickly programmed or written by tunneling. Flash EPROM has reached 4Mbit/chip density in 1990 from Intel production and still increasing, and is destined to replace the magnetic disks in notebook-size portable personal computers and other pocket size electronic equipment. Some of the devices using tunneling are described in chapter 6 on MOS transistors and MOS circuit building blocks for integrated circuits.

Single electron tunneling through a rectangular and triangular potential barrier in vacuum were described in sections 153 and 154 respectively. When the effective masses are used, these vacuum results are applicable to electron or hole tunneling through triangular potential barriers in semiconductor junction devices such as the p/n junction tunnel diode and the MOS capacitor just described. Later on, in section 385, these tunneling rates will be computed as a function of electric field and given in graphs.

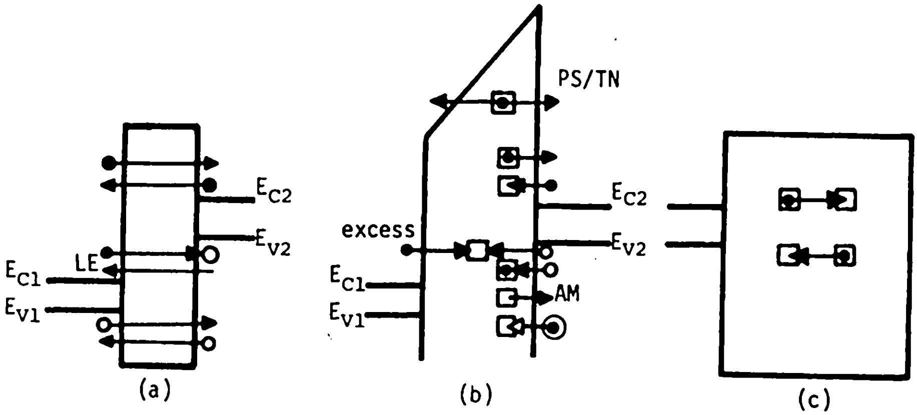


Fig.36n0 The transition energy band diagram of the three elastic tunneling mechanisms in a semiconductor/insulator structure. (a) Interband and inter-conduction-valence band tunneling. (b) Band-trap tunneling. (c) Intertrap tunneling.

36n4. Inelastic Tunneling

The electron energy change of the nine inelastic tunneling mechanisms (14, 24 and 34 in Table 360.1) is supplied or dissipated by phonon, photon, or another electron or hole. Some of the transitions of the nine mechanisms are shown in Figs.36n4(a)-(i). Because of the need of a third particle to supply or dissipate the electron energy change, the inelastic tunneling rate is smaller than the corresponding elastic tunneling transition. Nevertheless, they are responsible for the fine structures in the current-voltage characteristics of tunnel devices.

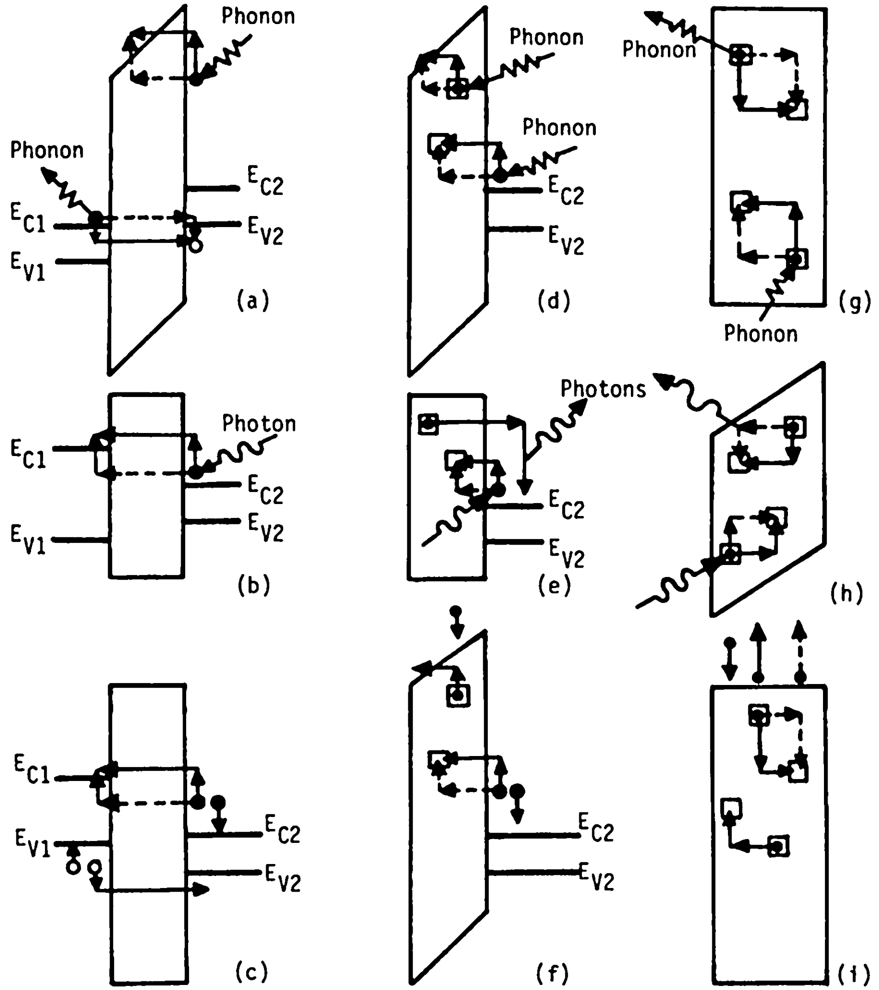

Phonon
Phonon
E_{C1}
E_{C2}
E_{V2}
E_{V1}
(a)
Phonon
Phonon
E_{C2}
E_{V2}
(d)
Phonon
Phonon
(g)
E_{C1}
Photon
E_{C2}
E_{V2}
E_{V1}
(b)
Photons
E_{C2}
E_{V2}
(e)
(h)
E_{C1}
E_{V1}
E_{C2}
E_{V2}
(c)
E_{C2}
E_{V2}
(f)
(i)

Fig.36n4 The transition energy band diagram of nine inelastic (thermal, optical, Auger-impact; and interband, band-trap, intertrap) tunneling mechanisms in semiconductor/insulator structures.

36n5 Collective Transitions

Collective energy exchange transitions are labeled 15, 25 and 35 in Table 360.1. The energy exchange involves and collective wave motion of all the valence electrons in a solid. The collective wave is quantized and the quanta of energy is known as the **plasmon**. Since the valence electron concentration is 10^{21}-10^{23} cm^{-3}, the energy of the plasmon in solids (proportional to $n_v^{-1/2}$) covers the range of 4eV to 30eV. This falls into the range of accelerating potential or voltage applied to semiconductor devices. Thus, one would expect plasmon creation and destruction to be important GRT mechanisms of electrons and holes during device operations at high voltages and under exposure to energetic particles (keV and MeV electrons, protons, and ions) and ionization radiations (keV x-rays and gamma rays).

Figures 36n5.1(a)-(d) give the transition energy band diagram of several GRT transitions with plasmon as the sole or the principal energy exchange mechanism. Figures (a) and (b) depict interband recombination and generation of an electron-hole pair involving a high energy electron or hole whose kinetic energy is greater than the plasmon energy. A plasmon is generated to dissipate the recombination energy in (a), and a plasmon is absorbed to supply the generation energy in (b). Figures (c) and (d) show two of the four interband recombination-generation processes via the Auger-impact transitions with a plasmon dissipates the recombination energy and supplies the generation energy. Figures (e) and (f) show two of the four band-trap capture-emission transitions in which the plasmon accounts for the electron energy changes.

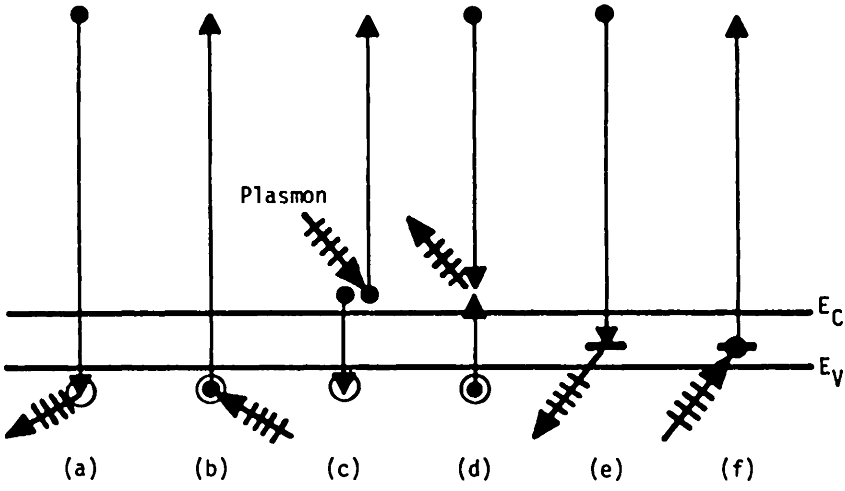


Fig.36n5.1 Transition energy band diagrams of the plasmon-assisted energy exchange mechanism. (a) and (b) Interband. (c) and (d) Interband Auger-impact. (e) and (f) Band-trap transitions.

370 Lifetimes

The net generation-recombination rates or generation-minus-recombination rate given by g_N-r_N and g_P-r_P in the first two Shockley Equations, (350.1) and (350.2), are usually written in terms of an electron lifetime and a hole lifetime, known as the **phenomenological lifetimes**. Their **operational definitions** are

$$r_N - g_N = (n - N_E)/\tau_n \equiv \delta n/\tau_n \tag{370.1}$$

and

$$r_P - g_P = (p - P_E)/\tau_p \equiv \delta p/\tau_p \tag{370.2}$$

Here, n=n(x,y,z,t) and p=p(x,y,z,t) are the electron and hole concentrations and they can be both space and time dependent. N_E and P_E are the equilibrium electron and hole concentrations which satisfy the mass action law, (242.2A),

$$N_E P_E = n_i^2. \tag{370.3}$$

The differential quantities designated by a leading lower-case delta, $\delta n \equiv n-N_E$ and $\delta p \equiv p-P_E$, are known as the **excess electron** and **excess hole concentrations** which can be negative indicating a **deficiency** when the electron and hole concentrations are decreased to below their equilibrium values by an applied voltage. Since carrier concentrations are particle concentrations, negative values tend to obscure the basic physics and seriously confuse a line of thoughts. However, the negative excess carrier concentration notation has been used in many textbooks and by many early authors and device pioneers.

Equations (370.1) and (370.2) **define** the electron and hole lifetimes, τ_n and τ_p, which are always positive quantities but may be dependent on position, time, as well as the electron and hole concentration variables. When they are constants and independent of the electron and hole concentration variables, the Shockley equations can be linearized and three characteristic lifetimes can be uniquely defined, one for electrons, one for holes, and a third for trapped electrons if a single-level electron trap is present, or for trapped holes if a single-level hole trap is present.

The electron and hole lifetimes can be related to the recombination and generation rate coefficients of each of the 18 transition processes given in Table 360.1. Analytical solutions for four simplest cases will be obtained to illustrate this dependency. These are the interband thermal and optical processes, 11 and 12 in Tables 360.1, and the band-trap thermal (Shockley-Read-Hall) and optical processes given by the elements 21 and 22.

371 Interband Thermal and Optical Recombination Lifetimes

In the interband thermal mechanism (process 11), the recombination rate, given by $r_N = r_P$, must be proportional to the concentration of electrons and holes,

$$r_{11} = r_N = r_P = r^t np. \tag{371.1}$$

The generation rate is independent of n and p if n and p are low, thus,

$$g_{11} = g_N = g_P = g^t \tag{371.2}$$

The two rates are combined and expanded using the equality $n=N_E+\delta n$ and $p=P_E+\delta p$ defined in (370.1) and (370.2), giving

$$r_N - g_N = r_P - g_P = r^t np - g^t \tag{371.3}$$
$$= r^t[(N_E+\delta n)(P_E+\delta p)] - g^t \tag{371.4}$$

We shall now try to eliminate either r^t or g^t from the above equation by using the thermal equilibrium condition, i.e., the thermal equilibrium generation rate must be equal to the thermal equilibrium recombination rate. Thus,

$$r_{NE} - g_{NE} = r_{PE} - g_{PE} = r_e{}^t \cdot N_E P_E - g_e{}^t = 0 \tag{371.5}$$

where subscripts 'e' and 'E' denote equilibrium. Thus, (371.5) connects $r_e{}^t$ to $g_e{}^t$ via

$$g_e{}^t = r_e{}^t \cdot N_E P_E = r_e{}^t \cdot n_i{}^2 \tag{371.6}$$

Next, a crucial assumption is made on the thermal generation and recombination coefficients, g^t and r^t. They are assumed to be constant and to remain at their equilibrium value when a force (electric field or light) is applied to the semiconductor, i.e., $r^t \simeq r_e{}^t$ and $g^t \simeq g_e{}^t$. Using this **small-deviation from thermal equilibrium** assumption and substituting (371.6) into (371.4),we have

$$\begin{aligned} r_N - g_N = r_P - g_P &= r^t[N_E P_E + N_E\delta p + P_E\delta n + \delta n\delta p] - g^t \\ &\simeq r_e{}^t[N_E P_E + N_E\delta p + P_E\delta n + \delta n\delta p] - r_e{}^t N_E P_E \\ &= r_e{}^t[N_E\delta p + P_E\delta n + \delta n\delta p] \end{aligned} \tag{371.7}$$

An equation relating δn to δp is still needed in order to obtain an explicit expression for the lifetimes. This is the charge neutrality condition discussed in section 242 which gives

$$\rho = q(p - n + N_{DD} - N_{AA} - n_T) = 0$$

or

$$\delta\rho = q(\delta p - \delta n - \delta n_T) = 0 \tag{371.8}$$
$$\delta p = \delta n + \delta n_T \simeq \delta n. \tag{371.8>}$$

In approximation (371.8A), the concentration of the recombination centers is assumed to be very small compared with the change of the electron and hole concentration, i.e., $\delta n_T << \delta n$ and $\delta n_T << \delta n$. This is known as **negligible trapping**. It may not be a good approximation in practical devices but it makes the analysis simple since we then have $\delta n=\delta p$ in (371.8A) which can be used in (371.7) to give

$$r_N - g_N = r_P - g_P \simeq r_e{}^t[N_E + P_E + \delta n]\delta n. \tag{371.9}$$

This is used to obtain an expression for the electron and hole lifetimes defined by (370.1) and (370.2). Since we have $\delta n=\delta p$ from the assumption of negligible trapping or negligible concentration of recombination center, (370.1) and (370.2) shows that **the electron and hole lifetimes are equal when trapping is negligible**. Using (371.9), then

$$\tau_n = \tau_p = \frac{1}{r_e^t(N_E+P_E+\delta n)} \tag{371.10}$$

$$= \frac{1}{r_e^t(N_E+P_E)} \quad \text{(Low Injection Level)} \tag{371.10A}$$

$$= \frac{1}{r_e^t(\delta n)} \quad \text{(High Injection Level).} \tag{371.10B}$$

To get the approximate result of (371.10A), we assume that the excess carrier concentration, δn and δp, are small compared with the dopant impurity concentration or the equilibrium majority carrier concentration, N_E or P_E, i.e. $\delta n = \delta p << $ (larger of N_E and P_E). This is known as the **low level injection condition** and the lifetimes are known as the **low level lifetimes**. The term 'injection', coined by Shockley, came from its use in forward biased emitter-base junction of bipolar transistors. In the opposite limit, we obtained the **high level lifetimes**, (371.10B), when $\delta n = \delta p >> $ (larger of N_E and P_E).

A similar analysis for the interband optical mechanisms (process 12) can also be made following exactly the same algebra just presented for the interband thermal mechanism (process 11). We have to only replace the superscript t (for thermal) by o (for optical) and need not redo the algebra. **This shortcut demonstrates the power from judicial choice of notation.**

These results show that the low-level recombination lifetime of electrons and holes due to the interband thermal and optical processes (processes 11 and 12 in Tables 360.1) are **inversely proportional** to (i) the thermal-equilibrium recombination rate coefficients, r_e^t or r_e^o, and (ii) the sum of equilibrium electron and hole concentrations, N_E+P_E. Usually, the majority carrier concentration is much larger than the minority carrier concentration so that the majority carrier concentration in the denominator, N_E+P_E, will dominate. Another important conclusion is that the lifetime at low injection level is independent of the excess carrier concentration or the 'injection' level. This is an expected result at low levels which **linearize** the problem. Otherwise the level assumed is not sufficiently low and the term δn (or δp) in the denominator of the lifetime expression (371.10) cannot be dropped and the error of (371.10A) would be significant.

At high injection levels when $\delta n >> N_E+P_E$, the lifetime given by (371.10B) is no longer dependent on the dopant impurity density or equilibrium carrier concentration but solely on the injected or excess carrier concentration. This shows that the lifetime is no longer a constant; the system is **nonlinear**; and the system cannot be characterized by a set of discrete and constant lifetimes or time constants.

372 Band-Trap Thermal (SRH) and Optical Recombination Lifetimes

The band-trap thermal and optical lifetimes (processes 21 and 22) are different from the interband lifetime given by (371.10). This is expected because interband recombination extends over the entire crystal (not localized) while band-trap recombination is localized at a trap. An explicit difference is that the band-trap lifetimes do not depend strongly on N_E and P_E if the semiconductor is extrinsic or not intrinsic and the injection level is low. The band-trap recombination lifetimes can be expediently derived without going through the detailed algebra normally given a traditional derivation which includes all of the four transition processes indicated in Figs.3611.1 and 3612.1.

As an example, consider the electron (minority carrier) lifetime in a p-type semiconductor containing an electron trap which is a deep-level GRT center that has an electron bound state. Let the total concentration of the electron trap be N_{TT} where T stands for trap and the repeated double subscript TT denotes the total number which will not change with a change in the applied voltage, an IEEE notation convention. In contrast, the single subscript in N_T, denotes the trapped electron concentration, which can change with the applied voltage. (See Appendix A at the end of this book.)

The electron capture rate is proportional to the product of the instantaneous electron concentration, n, and the unoccupied trap concentration, $p_T = N_{TT} - n_T$, where n_T is the instantaneous trapped electron concentration. Let the electron capture rate coefficient be denoted by $c_n{}^t$ and the emission rate coefficient of trapped electron be denoted by $e_n{}^t$, then the net rate of electron capture over trapped electron emission at the trap is

$$\begin{aligned} r_N - g_N &= c_N - e_N \\ &= c_n{}^t \cdot n \cdot p_T - e_n{}^t \cdot n_T = c_n{}^t \cdot n \cdot (N_{TT} - n_T) - e_n{}^t \cdot n_T. \qquad (372.1) \end{aligned}$$

But, $c_N - e_N = 0$ at thermal equilibrium. Thus, $c_{ne}{}^t \cdot N_E \cdot (N_{TT} - N_{TE}) = e_{ne}{}^t N_{TE}$. Again assume small deviation from thermal equilibrium so that the capture and emission rates, $c_n{}^t$ and $e_n{}^t$, are nearly constant. Then, their nonequilibrium values can be approximated by their equilibrium values, $c_n{}^t \approx c_{ne}{}^t$ and $e_{nt} \approx e_{ne}{}^t$. Two other approximations are also made as indicated in the following algebra to give the lifetime defined by (370.1). Let $n = N_E + \delta n$ and $n_T = N_{TE} + \delta n_T$.

$$\begin{aligned} r_N - g_N &= + c_n{}^t \cdot n \cdot (N_{TT} - n_T) \qquad\qquad - e_n{}^t \cdot n_T && (372.1) \\ &= + c_n{}^t (N_E + \delta n)(N_{TT} - N_{TE} - \delta n_T) - e_n{}^t (N_{TE} + \delta n_T) && (372.1A) \\ &= + c_n{}^t N_E (N_{TT} - N_{TE}) + c_n{}^t \delta n (N_{TT} - N_{TE}) - c_n{}^t N_E \delta n_T - c_n{}^t \delta n \delta n_T \\ &\quad - e_n{}^t N_{TE} - e_n{}^t \delta n_T \\ &\approx + c_{ne}{}^t N_E (N_{TT} - N_{TE}) + c_{ne}{}^t \delta n (N_{TT} - N_{TE}) - \ 0 \ - \ 0 \\ &\quad - e_{ne}{}^t N_{TE} - \ 0 \end{aligned}$$

$$= + c_{ne}^{t}\delta n(N_{TT}-N_{TE}) \qquad (372.1B)$$

$$= + c_{ne}^{t}(n - N_{E})(N_{TT}-N_{TE})$$

$$\approx + c_{ne}^{t}(n - N_{E})N_{TT} \quad (\text{if } N_{TT}>>N_{TE}) \qquad (372.1C)$$

$$\equiv + (n - N_{E})/\tau_{n}. \qquad (372.1D)$$

The three assumptions made in (372.1A) to give (372.1B) are: (i) the small-deviation from thermal equilibrium assumption on c_n^t and e_n^t, (ii) the small trap density assumption to drop the δn_T term, and (iii) the small-deviation from electrical equilibrium or low-level assumption to drop N_{TE} compared with N_{TT}. (iii) is increasingly more accurate as the semiconductor becomes more extrinsic or more p-type. Thus, the lifetime of electrons in extrinsic p-type semiconductors due to the Shockley-Read-Hall or SRH thermal capture-emission mechanism from (372.1C) and (372.1D) is

$$\tau_n = 1/(c_{ne}^{t}N_{TT}) \equiv \tau_{n0}. \qquad (372.3)$$

Similarly, the low level SRH lifetime of holes in extrinsic n-type semiconductors is

$$\tau_p = 1/(c_{pe}^{t}N_{TT}) \equiv \tau_{p0}. \qquad (372.4).$$

Assumption (372.1C) substantially simplifies the analysis as indicated above. Its key consequence is that all the electron traps are empty or not occupied by electrons in the p-type semiconductor, resulting in (372.3); or all filled or occupied by electrons in the n-type semiconductor, resulting in (372.4).

In spite of the small-deviation assumption, (iii), the final results are asymptotically correct. They are valid for either an electron or hole trap in either a n-type or p-type extrinsic semiconductor at low injection level or small deviation from electrical equilibrium as long as assumptions (i) and (ii) are also satisfied. The results show that the lifetimes are independent of the dopant impurity concentration or the equilibrium electron or hole concentration. However, the lifetimes increase towards $\tau_{n0}+\tau_{p0}$ if assumption (iii) becomes inaccurate, for examples, when the material becomes intrinsic or the injection level increases. The results, (372.3) and (372.4), show that the SRH lifetimes can be changed and designed by controlling the concentration of the trapping centers since the lifetimes are inversely proportional to N_{TT}.

A similar derivation can be made for the band-trap optical transitions. The optical lifetimes are obtained by replacing the superscript t with o. This again demonstrates the power of proper choice of notation.

373 Lifetimes for Simultaneous Presence of Many GRTT Mechanisms

When there are more than one energy exchange mechanism and when both interband and band-trap transition processes are present in a semiconductor, the recombination-generation-capture-emission rates will add and the reciprocal lifetimes of each mechanism will also add. Since there are more recombination 'channels' or paths, the lifetime will be lower. For example, if both the thermal and optical band-trap mechanisms are present on one recombination center species, (372.4) would become

$$\tau_n = 1/[(c_n{}^t + c_n{}^o)N_{TT}]. \tag{373.1}$$

380 PHYSICS AND DATA OF THE GRTT RATE COEFFICIENTS

The physics of the rate coefficients, such as r^t, g^t, r^o, g^o, $c_n{}^t$, $e_n{}^t$, $c_p{}^t$, $e_p{}^t$, $c_n{}^o$, $e_n{}^o$, $c_p{}^o$ and $e_p{}^o$, can be understood using simple particle models. The mathematical formulae relating the rates to the fundamental material constants of the semiconductor can be derived by solving the Schrödinger equation of the system of many particles. For example, the interband thermal generation rate of electron-hole pairs, g^t, can be related to and hence computed from the intrinsic (optical) absorption spectra of the semiconductor. Similarly, the interband thermal recombination rate of electron-hole pairs, r^t, can be computed from the black-body radiation spectrum of the material. The latter analysis was presented by Van Roosbrook and Shockley in 1953 who estimated an intrinsic lifetime of about 1 ms for Si. An estimate by later authors on GaAs gives 1 ns since GaAs has a direct energy gap. Similarly, the interband optical generation rate of electron-hole pairs at a given photon energy, $h\nu$, is directly given by the intrinsic (optical) absorption coefficient at $h\nu$ and the photon flux at $h\nu$. The interband optical recombination rate of electron-hole pairs gives the light emission efficiency of the material.

A more detailed but still very elementary description of the SRH thermal capture and emission rate coefficients, c^t and e^t, and the extrinsic optical emission rate, e^o, is given in the next two sections, 381 and 382. They are followed by a description of the interband optical rate, g^o, and impact generation rates, g^n and g^p, in sections 383 and 384. The interband tunneling rate is discussed in 385. Selected reliable experimental data are also given to illustrate the description of the basic physics.

381 Thermal (SRH) Capture and Emission Rates

The capture coefficients, c's, have the dimension of cm^3/sec. A simple physical picture can be developed for these rate coefficients. Consider a volume element $dV = dxdydz$ indicated in Fig.381.1 where the concentration of the recombination-generation-trapping center is N_{TT}. Each center is represented by a ball with radius 'a' or projected area of $\sigma_n{}^t = \pi a^2$ which defines a cylinder within which an electron is captured by the trap for certain. Then, the rate at which the electrons are captured by the centers in the volume dV is equal to the product of the electron velocity, θ_n, the number of recombination centers, $dV \cdot N_{TT}$, and the electron density, n,

$$(dV \cdot N_{TT}) \cdot \sigma_n{}^t \cdot \theta_n \cdot n. \qquad (381.1)$$

The volume density of the rate of capture of electrons is then

$$N_{TT} \cdot \sigma_{nt} \cdot \theta_n \cdot n \qquad (381.2)$$

which was written as $c_n{}^t \cdot n \cdot N_{TT}$ in (372.1)-(372.3). Thus,

$$c_n{}^t = \sigma_n{}^t \cdot \theta_n \qquad (381.3)$$

that is, the capture rate is the product of the capture cross section, $\sigma_n{}^t$, and the thermal velocity of the electrons, θ_n. This result can be readily generalized to many electrons with many different velocities or kinetic energies. Since the capture cross-section may vary with the electron velocity or kinetic energy, the above expression should be averaged over all the electrons in the conduction band. It is a macroscopic capture coefficient.

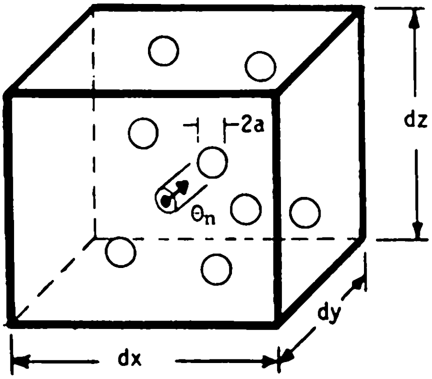


Fig.381.1 A differential volume element dxdydz of a semiconductor containing N_{TT}dxdydz number of recombination-generation-trapping centers. The centers are represented by balls to illustrate the significance of the capture rate coefficients, c_n and c_p.

The thermal emission rate, $e_n{}^t$, can also be understood by a similar particle model, namely, $e_n{}^t \cdot (N_{TT} \cdot dV)$ is obviously the rate of emission of electrons from the $(N_{TT} \cdot dV)$ traps in the volume element dV if all the traps are each occupied by an electron.

The thermal equilibrium emission rate can be calculated from the thermal capture rate or vise versa. Their relationship is obtained by setting (372.1) to zero at equilibrium to give $e_{ne}{}^t N_{TE} = c_{ne}{}^t N_E (N_{TT} - N_{TE})$. At equilibrium, the electron trap occupation factor, has an appearance similar to the Fermi function for the electron energy levels in the band: $f_{TE}(E_T) = N_{TE}/N_{TT} = \{1 + \exp[(E_F - E_T)/kT]\}$, but it is fundamentally different from the Fermi function as discussed in section 252. E_T is the effective level which takes into account configuration and spin degeneracies. Thus,

$$\begin{aligned} e_{ne}{}^{t} &= c_{ne}{}^{t}\cdot N_E\cdot(N_{TT}-N_{TE})/N_{TE} \\ &= c_{ne}{}^{t}\cdot N_C\exp[(E_F-E_C)/kT]\cdot\exp[-(E_F-E_T)/kT] \\ &= c_{ne}{}^{t}\cdot N_C\exp[-(E_C-E_T)/kT] = c_{ne}{}^{t}n_1. \qquad (381.4) \end{aligned}$$

The density parameter, n_1, was first introduced by Shockley. It is defined by

$$\begin{aligned} n_1 &= N_C\exp[-(E_C-E_T)/kT] \qquad (381.5A) \\ &= n_i\exp[+(E_I-E_T)/kT] \qquad (381.5B) \end{aligned}$$

which shows that n_1 is the electron concentration when E_F coincides E_T. It is also the equilibrium reaction constant if the electron capture and emission transitions at a trap are treated as the reversible chemical reactions: $n+(N_{TT}-n_T) \longleftrightarrow n_T$.

An order of magnitude estimate of $c_n{}^t$ can be made using a sequence of simple physical considerations. θ_n is the thermal velocity of the electrons which is about 10^7 cm/sec at 300°K computed from $m\theta^2/2=3kT/2$. The capture cross section is roughly given by πa^2 where a is the radius of the trapping center or the radius of the trap potential well. Assume $a=6A$ which is about a lattice constant. Then, a recombination volume of a^3 would contain one trapping center surrounded by about 20 host atoms. The number 20 is consistent with energy conservation and exchange. When the electron is captured by a deep-level trap in Si whose energy level is near the midgap, $E_C-E_T \approx E_G/2$, the capture energy transferred to the surrounding host atoms is about 500 meV for ($E_{G\text{-}Si}/2=556$ meV). To dissipate this capture energy, the 20 Si host atoms surrounding the trapping center in the recombination volume must increase their vibration frequency since their vibration amplitude cannot be more than about one tenth of the interatomic spacing, about a/5 in the diamond lattice. Otherwise, local structural transformation or local melting would occur in the recombination volume and the recombination impurity that gives the electronic trap would move to the surface and cannot stay inside the Si crystal. This does not seem to happen in practice for many recombination impurities in Si. Increasing the vibration frequency is also limited since the average phonon energy in Si and most solids is only about kT or 25 meV. There are two possible ways to dissipate the capture energy of 500 meV at a midgap trap in Si: (i) one of the surrounding host atom or the impurity atom itself will vibrate 20 times to sequentially emit 20 phonons (20x25meV=500meV) and (ii) 20 host atoms surrounding the trapping center will each emit a phonon of about 25 meV. Obviously the latter has a higher quantum mechanical transition probability. This is a qualitative picture but it does bring out the physics which is helpful if a detailed quantum mechanical calculation of the capture rate is made. Such a picture depicted in (ii) has not been used in several theoretical attempts in the past. From the assumed $a=6A$, then

$$\sigma_n{}^t = \pi a^2 = 10^{-14}\ \mathrm{cm}^2 \qquad (381.5)$$

and

$$c_n = \sigma_n \theta_n = 10^{-14} \times 10^{7} = 10^{-7}\ \mathrm{cm}^3/\mathrm{sec}. \qquad (381.6)$$

If $N_{TT} = 10^{14}/\mathrm{cm}^3$ or one part per billion of Si atom (10^{23} Si/cm^3) then

$$\tau_{n0} = 1/(c_n N_{TT}) = 1/(10^{-7} \times 10^{14}) = 10^{-7}\ \mathrm{sec}. \qquad (381.7)$$

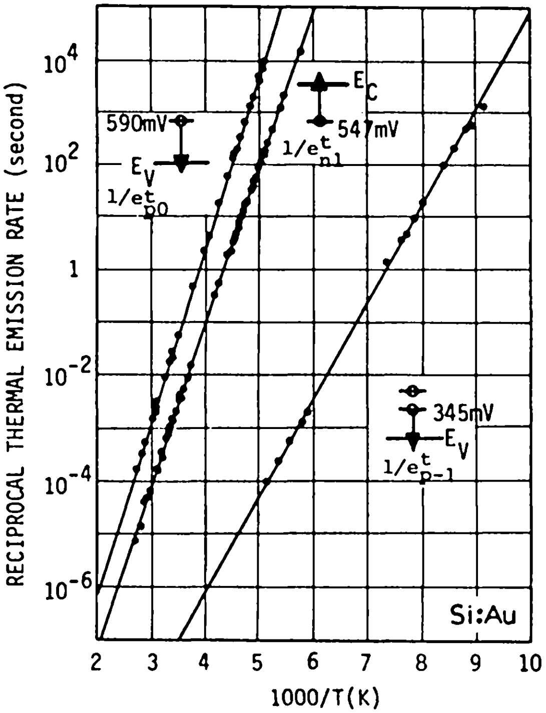


Fig.381.2 Experimental data of the thermal emission rate of electrons and holes from the two gold recombination-generation-trapping centers in Si. [From C.T.Sah, L.Forbes, L.L.Rosier, A.F.Tasch,Jr., and A.B.Tole, Applied Physics Letters, 15, 161-164, 15 September 1969.]

The thermal emission rate can also be estimated. For a trap energy level at the midgap or the intrinsic Fermi level, $E_T \approx E_I$, we have $n_I \approx n_i = 10^{10}\mathrm{cm}^{-3}$. Thus, $e_{ne}{}^t = c_{ne}{}^t n_I \approx 10^{-7}(\mathrm{cm}^3/\mathrm{s}) \times 10^{10}(1/\mathrm{cm}^3) = 1000\mathrm{s}^{-1}$ or 1000 trapped electrons are emitted or released per second. It is evident that the thermal emission rate is highly dependent on temperature. It is thermally activated, i.e., it takes thermal energy to emit or release the trapped electron. The thermal activation energy is equal to E_C-E_T for trapped electrons or E_T-E_V for trapped holes.

Experimental data of the thermal capture rates of electrons and holes at the recombination centers in semiconductors are few and inaccurate due to the difficulty of measuring the very low concentration of the recombination centers although the lifetimes can be accurately measured. The thermal emission rates, such as e_n^t in (372.1), are easier to measure. The most accurate set is that of the gold center in Si since gold has been very successfully used to increase the switching speed of Si diodes and transistors. Gold has two energy levels in the silicon energy gap, one near the midgap and the other in the lower half gap. Both are hole traps and efficient recombination levels for minority carriers. The midgap trap is also a very efficient generation level. The thermal emission rate of electrons at the midgap level and of holes at both the midgap and the lower-gap levels have been measured by this author and his graduate students during 1967-1969 as a function of temperature covering nearly ten decades of rate values. This is shown in Fig.381.2.

382 Optical Emission Rate

Optical emission rate of trapped electrons is given by $e_n^o \cdot N_T$ where N_T is the concentration of the optical center that is occupied by electron. In an extrinsic n-type semiconductor, most of the optical centers are occupied by electron if it is deep. Thus, N_T can be approximated by the total concentration of the optical center N_{TT} (occupied plus unoccupied).

The traditional expression of the photo-ionization rate of an optical center in a solid or semiconductor is given by $\sigma_n^o \cdot N_T \cdot \Phi^o = \alpha_n^o \cdot \Phi^o$. Here Φ^o is the photon flux (in photon/cm^2-sec). σ_n^o (in cm^2) is the photo-ionization cross-section of an occupied (neutral or charged) optical center. It is also the photoexcitation cross-section of the trapped electron at either a neutral or charged optical center. α_n^o (in cm^{-1}) is the traditional **extrinsic absorption coefficient.** Equating these expressions,

$$e_n^o \cdot N_T = \sigma_n^o \cdot \Phi^o \cdot N_T = \alpha_n^o \cdot \Phi^o, \tag{382.1}$$

then the optical emission rate of a trapped electron is given by

$$e_n^o = \sigma_n^o \cdot \Phi^o. \tag{382.2}$$

It is proportional to the fundamental parameter, σ_n^o (the photoexcitation cross-section), and independent of the concentration of the optical center, N_{TT}. The conventional absorption coefficient, α_n^o, is a function of the concentration of the optical center N_T, thus, not a fundamental parameter.

383 Interband Optical Generation Rate

The interband optical or radiative recombination of electron-hole pairs with the emission of photons was described in section 3612. The light emission of this process is the basis of the light emitting p/n junction diode and injection p/n junction diode laser. The inverse process, electron-hole pair generation, involves the absorption of light with photon energy greater than the energy gap via breaking the covalent bond. This is commonly known as **intrinsic absorption** and it is the fundamental mechanism underlying the operation of photo detectors and solar cells. The volume generation rate of electron-hole pairs, denoted by $g_{21} = g_N{}^o = g_P{}^o = g^o$ (pair/cm^3-s) in (3612.1) is directly related to the intrinsic absorption coefficient of the material. The relationship is

$$g_n{}^o = g_P{}^o = g^o = \alpha^o \Phi^o. \qquad (383.1)$$

α^o is the intrinsic absorption coefficient in the unit of number of photons absorbed per thickness (cm^{-1}). Φ^o is the photon flux (photon/cm^2-sec). Data of the absorption coefficient as a function of wavelength has been obtained for many materials. Data for Ge, Si, GaAs, Ge and CdS are shown in Fig.383.1

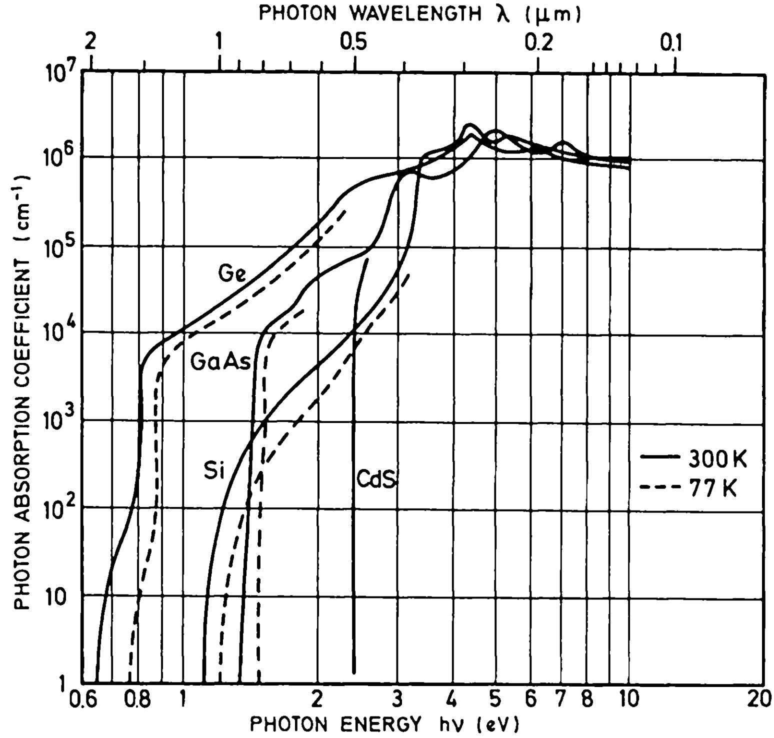


Fig.383.1 The intrinsic optical absorption coefficient, α_0, of Ge, S, GaAs, and CdS as a function of photon energy. The solid curves are 300K data and the dashed curves are 77K data.

384 Interband Impact Generation Rate

The interband impact generation of electron-hole pairs by energetic electrons or holes and its inverse, the Auger recombination process, were described in section 3613. Its practical importance includes the following. (i) It determines the current-breakdown voltage of a p/n junction which is used as a highly accurate and stable voltage reference. (ii) The large multiplication of electron-hole pairs in the high electric field of a p/n junction prior to current breakdown has been used to multiply the photo-generated electron-hole pairs to give tremendous optical gain. (iii) Impact generated high-energy electron-hole pairs in the high-field surface space-charge layer is the basic mechanism of charge injection into the floating gate of the MOS memory transistors to write a bit in UV-EPROM chip (Ultraviolet Light Erasable Programmable Read Only Memory) which has been manufactured at 4Mbit/chip density in 1990.

The interband impact generation rates due to energetic electron and energetic hole impact, defined by (3613.1) and (3613.2), are measured for many semiconductors as a function of the applied electric field that accelerates the electrons and holes. The data are obtained by measuring the multiplication factor of a current injected into a reverse biased p/n junction diode. The current is injected either by interband optical generation just described or by a forward biased emitter junction using a bipolar junction transistor structure. What are measured are the interband or pair generation **coefficients** due to electron and hole impact, known as the interband **impact ionization coefficients** and denoted by α^n and α^p. They are related to the interband generation rates, g^n and g^p defined in (3613.1) and (3613.2), by the following relationships.

$$g^n = \langle \alpha_n v_n \rangle \approx \alpha_n \theta_n \tag{384.1}$$

$$g^p = \langle \alpha_p v_p \rangle \approx \alpha_p \theta_p . \tag{384.2}$$

The interband impact ionization coefficients, α_n and α_p, are the number of electron-hole pairs generated per unit distance traveled by an energetic electron or hole. θ_n and θ_p are the scattering-limited velocities of the impacting electrons and holes and they are roughly equal to the thermal velocities, 10^7 cm/s, as indicated in subsection 314 and Fig.314.1.

Data on the interband impact ionization coefficients as a function of electric field are given in Fig.384.1 for Ge, Si, GaAs and GaP. They are nearly straight lines represented by

$$\alpha = \alpha_0 \exp(E_0/E). \tag{384.3}$$

This simple relationship is not empirical but has a firm theoretical basis developed by Shockley in 1961 known as the ballistic model. The characteristic electric field, E_1, is equal to the optical phonon energy divided by the electron (or hole) mean-free-path due to optical phonon scattering, $E_0 = \hbar\omega_0/q\lambda_0$. When the electron (or hole) gains substantial kinetic energy from the accelerating electric field, the rate of its energy loss increases since it does not have sufficient energy to emit or generate optical phonons which have higher energies than the acoustical phonons. Thus, electron scattering via optical phonon emission reduces the probability that the electron can be accelerated to the impact ionization threshold energy determined in section 3613. This reduction is given by the exponential factor in (384.3). If the density of the high energy electron is high, the electron energy distribution must be considered instead of considering only a few electrons because of the high electron-electron collisions owing to many electrons. Then, ionization rate is given by $\alpha=\alpha_2\exp(E_2/E)^2$. This is known as the diffusion model. The curvature is discernible in Fig.384.1.

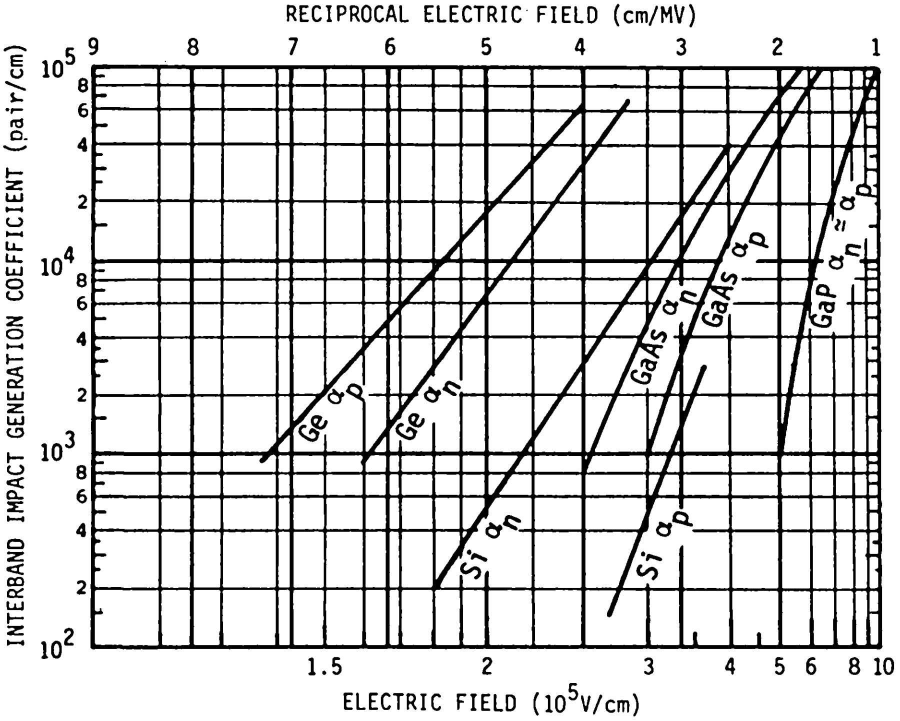


Fig.384.1 The interband impact generation coefficients or impact ionization coefficients by electron impact and by hole impact in Ge, Si, GaAs and GaP.

385 Interband Tunneling Rate

Tunneling transitions and their practical importance in devices were described in sections 36n0 and 36n4. The most important application today is the elastic tunneling of electrons and holes across the triangular potential barrier at the oxide/silicon interface of an oxidized silicon in MOS transistors. This is known as

Fowler-Nordheim tunneling. It is the mechanism in the operation of EEPROM memory transistor and the fundamental limitation of submicron Si MOS transistors and integrated circuits as the dimension decreases to below 0.5 micron.

The tunneling current is given by the product of the electron or hole velocity, the electron or hole concentration, and the tunneling probability across the triangular potential barrier at the SiO_2/Si interface. Using the tunneling probability through a triangular barrier given by (154.1A) or (154.1B), the electron and hole tunneling currents are

$$J_N = 2.6\times10^6 E_0^2 \exp(-238.5/E_0) \quad \text{A/cm}^2 \quad \langle 100\rangle \tag{385.1A}$$

$$J_N = 1.8\times10^6 E_0^2 \exp(-285.0/E_0) \quad \text{A/cm}^2 \quad \langle 111\rangle \tag{385.1B}$$

and

$$J_P = 2.\ \times10^6 E_0^2 \exp(-440.\ /E_0) \quad \text{A/cm}^2 \quad \langle \text{ave}\rangle \tag{385.2}$$

where E_0 (in MV/cm or 10^6V/cm) is the oxide electric field at the SiO_2/Si interface. The numerical values in the above formula were obtained by fitting the theory to the experimental data of Ziv Weinberg. The barrier height at the SiO_2/Si interface is 3.08±0.13eV for electrons and 4.3±0.6eV for holes. The orientation dependence for electrons arises from the orientation dependence of the 2-d density-of-state effective mass because the constant energy surfaces of the Si conduction band are ellipsoids and and not spheroids. The three equations are plotted in Fig.385.1. Note the three application specific scales on the right: pA/μm^2 for fundamental experiments, e/μs-μm^2 for memory write, and e/yr-μm^2 for 10-year operating life.

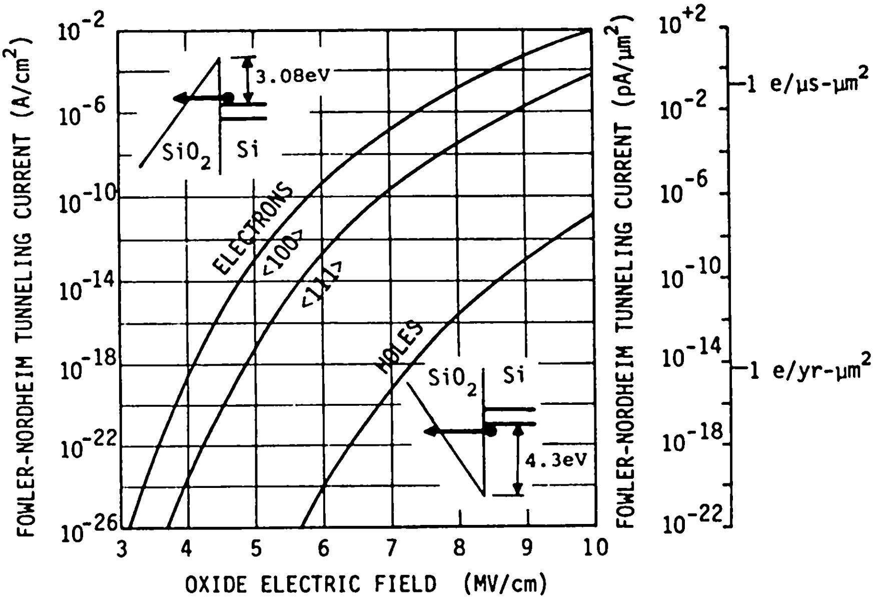


Fig.385.1 The Fowler-Nordheim tunneling currents through the SiO_2/Si interface.

386 Band-Trap Tunneling Rate

The elastic band-to-trap and trap-to-band tunneling transitions were described in section 36n0. They are the principal transition that gives the excess current in the negative resistance Esaki p/n junction diode. They are also responsible for charging and discharging the traps in the gate oxide film which causes the MOS transistor to drift or age, and the MOS circuits to fail. A trap potential well in a triangle oxide potential barrier, shown in Fig.386.1, is a representative model. The tunneling rate is given by

$$\begin{aligned}\omega &= (\pi^2/h^3)\surd(m_y m_z)(F_t/E_t)W^2\exp\{-(8\pi/3)[\surd(2m_x)]E_t{}^{3/2}/hF_t\}\\ &= 7.93264\text{x}10^{10}(F_t/E_t)\exp\{-68.28995E_t{}^{3/2}/F_t\}\ \text{s}^{-1}.\end{aligned} \tag{386.1}$$

F_t(MV/cm) is the constant oxide electric field. E_t(eV) is trap depth. The exponential dependence can be readily obtained from the simple barrier tunneling formula, (154.1). The numerical result assumes $m_x = m_y = m_z = m$ (free electron mass) and $W^2 = 10^{-24}V^2cm^3$ for the tunneling transition probability. The current from elastic tunneling of Si electrons into an oxide trap can be estimated from

$$J \approx 10^{-14}(N_{TT}-N_T)x_T\exp\{-(240/E_0)[1-(\Phi_T/\Phi_B)^{3/2}]\}\ \text{A/cm}^2. \tag{386.2}$$

$N_{TT} - N_T$ is the concentration of the unoccupied trap. As illustrated in Fig.386.1, x_T is trap distance from the oxide/Si interface, Φ_T is the trap depth, Φ_B is the barrier height, and E_0 (MV/cm) is the oxide field. The current from a trapped electron tunneling out of an oxide trap into the SiO_2 conduction band is

$$J \approx 10^{-14}(N_T)x_T\exp\{-(240/E_0)(\Phi_T/\Phi_B)^{3/2}\}\ \text{A/cm}^2. \tag{386.3}$$

For holes, the characteristics electric field field of the SiO_2/Si interfacial potential barrier for electron tunneling, 240MV/cm, is replaced by 440MV/cm because the oxide/Si barrier height for holes is 4.7eV compared with 3.1eV for electrons, giving the factor $(4.7/3.1)^{3/2} = 1.85$ or 1.85x240=440.

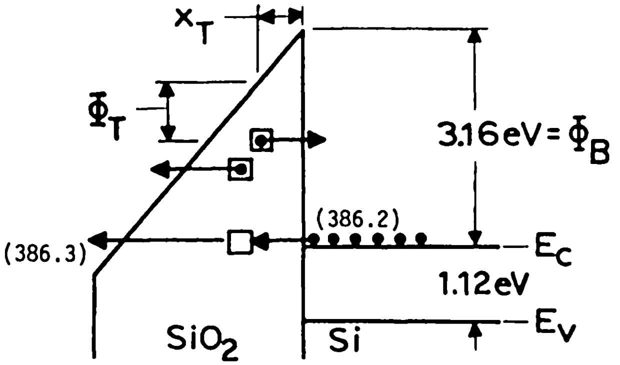


Fig.386.1 Oxide trap potential well and barrier heights for tunneling into and out of the trap.

399 BIBLIOGRAPHY

There are many text and reference books that describe and derive the continuity and current equations but only one that covers some of the generation-recombination-trapping-tunneling kinetics. Almost all textbooks have used the steady-state lifetimes in transient problems without justification. Thus, for further reading on the GRTT kinetics, original and review journal articles should be consulted, some of which are listed.

[399.1] William Shockley (Bell Telephone Laboratories, Shockley Transistor Corporation of Beckman Instruments, Inc., Stanford), **Electrons and Holes in Semiconductors**, D. Van Nostrand Company, Inc. New York, 1950. Drift current with ion and phonon scattering are derived in chapter 11 using the steady-state Boltzmann equation (p.274); diffusion current and continuity equation using phenomenological lifetime approximation are given in chapter 12.

[399.2] R.A.Smith (Edinburgh and MIT), **Semiconductors**, 2nd ed., Cambridge University Press, London, 1978. Chapters 5 through 9 (about 200 pages) describe the derivation using Boltzmann equation and applications of the continuity and current equations.

[399.3] Arthur C. Smith, James F. Janak and Richard B. Adler (MIT), **Electronic Conduction in Solids**, McGraw-Hill Book Company, New York, 1967. Appendix K gives the first compact derivation of the equations of semiconductor physics or the first five Shockley equations. Does not give the sixth Shockley equation.

[399.4] F. Reif (Berkeley), **Fundamentals of Statistical and Thermal Physics**, McGraw-Hill Book Company, New York, 1965. Chapters 13 and 14 give detailed classical derivation of the Boltzmann transport equations and the equilibrium and near-equilibrium solutions including current-charge continuity, and conservation of momentum and energy, which can be used for the electron and hole quantum quasi-particles.

[399.5] K. Seeger (Vienna and Boltzmann Institute), **Semiconductor Physics**, An Introduction, 2nd ed., Springer-Verlag, Berlin, 1982. Contains many examples with experimental data of the nearly equilibrium and hot electron transport properties in semiconductors.

[399.6] J.S.Blakemore (Honeywell, Florida Atlanta, Oregon Graduate Research Center), **Semiconductor Statistics**, Pergamon Press, New York, 1962. The only book gives comprehensive treatment of recombination kinetics.

[399.7] W.Shockley and W.T.Read,Jr. "Statistics of recombination of holes and electrons," Physical Review 87(9), pp.835-842, Sept.1952.

[399.8] Chih-Tang Sah and W.Shockley, "Electron-hole recombination statistics in semiconductors through flaws with many charge conditions," Physical Review 109(40, pp.1103-1115, Feb.15,1958.

[399.9] C.T.Sah, "The equivalent circuit model in solid-state electronics - Part 1: The single energy level defect centers," Proc.IEEE, 55(5), pp.654-771,May,1967; "... Part II: The multiple energy level impurity centers," pp.672-684; "...-III (Conduction and displacement currents)," Solid-State Electronics 13(12), pp.1547-1575, Dec.1970; "The equivalent circuit model in semiconductor transport for thermal, optical, auger-impact and tunnelling recombination-generation-trapping processes," Physica Status Solidi (a)7, pp.541-559, Oct.16,1971; "New integral representation of circuit models and elements for the circuit technique for semiconductor device analysis," Solid-State Electronics 30(12), pp.1277-1281, Dec.1987.

399 PROBLEMS

P310.1 Derive 3-d drift current density as a function of drift velocity.

P311.1 Verify that the rms noise in the drift velocity is $v_d(\text{noise}) = (q/m)E_x\surd[<\tau_f^2>-\tau^2]$. Show that because it is much smaller than the drift velocity, $v_d = (q/m)E_x\tau$, hence, it is even more smaller than the thermal velocity, $v_{th} = \surd<v_0^2> = (\surd 3kT/m)$.

P312.1 Describe the conditions at which the conductivity is an equilibrium parameter and hence a fundamental property of a solid.

P312.2 Is the definition of conductivity limited to thermal equilibrium?

P312.3 Is the definition of conductivity limited to a homogeneous solid?

P312.4 Is the drift current formulae limited to a homogeneous solid?

P312.5 Is the drift current formulae an equilibrium or nonequilibrium expression?

P312.6 A current of 1 A/cm^2 is forced through an n-type Si bar with a resistivity of 1 ohm-cm at 300K. What is the drift velocity of the electrons? How long does it take for an electron to drift through an Si bar of 1mm long?

P312.7 If the current density passing through the n-type 1 ohm-cm Si bar is increased to 10_4 A/cm^2 and 10^5 A/cm^2 (by reducing the cross-sectional area, what are the drift times for an electron across a 1mm long Si bar?

P312.8 The mobility of electrons in a certain n-type Si is about 1000cm^2/V-s. What is the scattering mean-free-time of the electrons? ($m_n = m$)

P312.9 Demonstrate that the condition of averaging over many scattering events is met using the data and results of the above three problems.

P312.10 From the drift velocity noise formulae, show that mobility fluctuation noise is not important compared with the thermal noise. Note that the mobility fluctuation noise has a white-noise (frequency independent) spectra up to a frequency of $1/2\pi\tau$ then drop off as f^{-2} just like the thermal noise.

P313.1 Drive the mobility formulae and its temperature dependence for a neutral scatterer whose scattering cross section is πa_s^2 and constant.

P313.2 Sketch the transverse acoustical lattice vibration wave at $t=0$ and a later time $t=t_1$ propagating on a one-dimensional chain with a wavelength of 8a.

P313.3 Sketch to scale the longitudinal and transverse displacement waves on a monoatomic one-dimensional lattice at a wavelength of $\lambda=4a$. Locate the vibration frequency of these two acoustical modes on the ω_q-q diagram such as that given by Fig.313.1(d).

P313.4 Sketch the phonon spectra of a tertiary compound semiconductor whose primitive unit cell contains three atoms such as $Ga_xAl_yAs_z$?

P313.5 Show that the scattering probability or scattering cross-section of electrons by acoustical phonon is proportional to kT by relating the rms displacement amplitude of the vibrating atoms to their average kinetic energy.

P313.6 Should the cross-sectional area of optical phonon scattering of electrons appear in (313.10) and why?

P313.7 Give the reason why the ionized impurity concentration in the mobility vs impurity concentration figure, Fig.313.5, is the sum of the ionized shallow donor and acceptor concentrations and not the difference.

P313.8 How are the mobility vs ion concentration curves in Fig.313.5 changed if a fraction of the ionized impurities becomes neutral?

P313.9 Obtain the numerical constants, μ_{I0}, for the ionized impurity scattering mobility formulae of electrons and of holes listed in Table 313.1 using the following information: (i) the two experimental mobility versus impurity ion concentration curves given in Fig.313.5, (ii) the empirically fitted formula of the lattice scattering mobilities given in Table 313.1, and (iii) the Matthiessen rule, (313.12).

P313.10 An Si crystal is doped uniformly with $1.000\text{x}10^{17}$ cm^{-3} of boron and $1.001\text{x}10^{17}$ cm^{-3} of arsenic. Compute the conductivity using Fig.313.5 or the formulae in Table 313.1.

P313.11 Matthiessen's rule given by (313.3) indicates that the reciprocal of the conductivities or the resistivity from each electron scatttering mechanism adds. But the hole and electron conductivities add to give the total conductivity and not their reciprocals, as indicated by (315.1) and (315.2) using (312.7A) and (312.7B). Explain. (Hint: two parallel channels, an electron and a hole.)

P314.1 Obtain the drift velocity saturation parameter, γ_n and γ_p, for hot electron and hot holes in silicon by fitting the empirical formulae given in section 314 to the drift velocity vs electric field data given in Fig.314.1.

P314.2 Calculate the critical electric field for electron drift velocity saturation, E_c, in Si at 300K using either the solution of Problem 31n.16 or from Fig.314.1 directly. Calculate E_c also for holes. How does the critical field change with temperature? Give the numerical values at 77K for pure Si. Does the saturated velocity vary with temperature and why?

P314.3 How do the saturated drift velocity value and the critical field for velocity saturation vary with dopant impurity concentration? Give the fundamental reasons.

P314.4 At what electric field will the drift velocity noise become comparable with the thermal noise? Assume the rms fluctuation of the free time is (i) 1% and (ii) 100% of the mean free time. Which is a better assumption and what is the physical basis.

P315.1 At what electron and hole concentrations is the resistivity of a semiconductor a maximum? Give numerical values for Si at 300K. Note in practice one cannot refine Si to such a purity, however, this can occur in a p/n junction where $N_{DD}=N_{AA}$ and $p=n=n_i$ at the intrinsic or p/n boundary. However, the resistivity is an irrelevant quantity in the space-charge layer of a p/n junction.

P315.2 Show that it is possible to have a higher electron conductivity than hole conductivity in a p-type Si. P-type is defined by $P>N$. This is again impractical and gives false impression for the same reason as above.

P315.3 A silicon is to be doped homogeneously by donor and acceptor impurities to make the electron and hole conductivities equal at room temperature which never happens in practice. Should the Si be doped with a group-III acceptor or group-V donor? Determine the required concentration of the donor and the acceptor impurities, assuming that all the impurities are ionized, i.e., each donor gives one electron and each acceptor gives one hole. Use the mobility curves in Fig.313.5 or formulae in Table 313.1.

P315.4 A state-of-the-art n/p/n Si transistor has a signal amplifying or electrically active p-type base layer whose volume is $1\mu m^3$ and which is doped with 10^{18} boron/cm^3. Suppose that the transistor will not work due to short-circuit paths in the base layer if the boron concentration in the p-base dropped below 10^{17}cm^{-3}. How many transistors must one fabricate to assure 100 good transistors? This is a real world problem.

P320.1 Show that if the electron mean-free-path is λ, then the probability that an electron will not be scattered while travelling through a distance x is $\exp(-x/\lambda)$.

P320.2 Estimate the mobility and diffusivity using (312.4B) and (320.8) for gas molecules at the standard condition, 300K and 760mm-Hg. You may look up the molecule density. Assume the molecules are hard spheres and $\tau_f=d/v_{th}=$constant where d is the intermolecular distance computed from the density and v_{th} is the thermal velocity of the molecules $\sqrt{kT/3M}$ where M is the mass of a molecule. What is the significance of the mobility if the molecules are all electrical neutral?

P321.1 Starting with the Einstein relationship, derive the Boltzmann relationship for the equilibrium electron concentration, $N(x)=N(x_1)\exp\{q[V(x)-V(x^1)]/kT]$. Is this result consistent with the fact negatively charged electrons concentrate more at locations of higher positive potential? Is it consistent with higher particle concentration in locations with lower potential energy?

P322.1 An n-type Si slice of a thickness L is inhomogeneously doped with phosphorus donor whose concentration profile is given by $N_{DD}(x) = N_0 + (N_L-N_0)(x/L)$. What is the formulae for the electric potential difference between the front and the back surfaces when the sample is at thermal and electric equilibria regardless of how the mobility and diffusivity varies with position? What is the formulae for the equilibrium electric field at a plane x from the front surface for a constant diffusivity and mobility? How is the formulae modified if diffusivity and mobility vary with impurity ion concentration?

P322.2 Derive the expression that gives the equilibrium electric potential as a function of position in the n-type Si slice given in Problem 322.1, $V_I(x)$.

P322.3 Derive the formulae of the drift and diffusion current densities of electrons and holes at the front and back surface of the Si slice given in problem P321.2 in terms of N_0, N_1 and L.

P322.4 If $N_0=10^{19}$cm^{-3}, $N_L=10^{16}$cm^{-3}, and $L=10^{-3}$cm$=10\mu$m, what are the numerical values of the quantities asked for in problems 321.2 and 321. Assume constant mobilities first. Then, take into account mobility variations with N_{DD} given in Fig.313.3.

P330.1 Show that the Fermi energy or Fermi potential is spatially constant in the n-type Si slice with linearly varying donor dopant impurity concentration as given in problem P322.1.

P330.2 The equilibrium electric potential variation in an n-type Si slice of thickness L is found to be given by $V_I(x)=V_0 + (V_L-V_0)(x/L)$. What are the equilibrium concentrations of electrons and holes as a function of x? Obtain the expressions of the drift and diffusion current densities as a function of x.

P330.3 Let $V_0 = 0.8$V, $V_L=0.4$V, and $L=10\mu m$, obtain the numerical values of the parameters asked for in problem P330.2.

P330.4 In principle, the equilibrium electric potential variation in problem P330.2 and the equilibrium electric potential difference in problem P322.1 cannot be measured? Why? How can they be measured approximately?

P330.5 Prove that $\partial E_F/\partial t=0$.

P331.1 A current of 1 A is forced through a homogeneous n-type 1 ohm-cm Si bar of cross-sectional area of $1mm^2$. What is the quasi-Fermi electric field of electrons in the Si bar? What is the quasi-Fermi electric field of holes in the Si bar? What are the electron and hole quasi-Fermi potential differences between the two ends of the bar of 1cm? What is the applied voltage across the 1cm bar? Are your answers consistent?

P331.2 The quasi-Fermi potentials are measurable quantities, why?

P331.3 Show that the product of the nonequilibrium electron and hole concentrations, NP, given by (331.7) and (331.10), reduces properly to the equilibrium expression, n_i^2.

P331.4 An inhomogeneously doped n-type semiconductor is at thermal and electrical equilibria. The donor concentration is given by $N_{DD}(x)=A\cdot\exp(-ax)$ where A and a are constants. The variation of N_{DD} with position is slow so that the semiconductor can be considered electrically neutral. Find the expression of the electric potential, V(x).

P331.5 An inhomogeneously doped Si sample is at thermal and electrical equilibrium. By some nonequilibrium measurements at T=290.14K, it was determined that this Si sample contains an equilibrium internal electric field of 10,000V/cm across a thin layer of 1.0μm thick inside the sample. What are the electron and hole concentration ratios on the two surfaces of the thin layer, N_1/N_2 and P_1/P_2, to account for this built-in electric field. Use kT/q=0.0250V at T=290.14K.

P331.6 If the above sample is n-type and the electron concentration at the location x_1 is $N(x_1)=10^{16}cm^{-3}$, what is the drift current density in A/cm^2? What is the concentration gradient at this point, $dN(x_1)/dx$?

P331.7 Define and derive the quasi-Fermi energy and potential for a species of ions.

P331.8 Define and derive the quasi-Fermi energy and potential for a species of neutral atoms.

P340.1 Show by elementary integration that the net steady-state volume generation-recombination rate of electrons (or holes) in a volume element dxdydz is given by the net electron current flowing out of the volume element.

P340.2 Given $V_N(x)=V_N(0) + [V_N(L)-V_N(0)](x/L)^2$ which is measured in an n-type Si sample, what are the electron and hole generation-recombination rates as a function of x to give this current in the 1-d n-Si bar in problem 331.1?

P350.1 Why do we not also have another equation for trapped holes similar to (350.6)? When do we need one?

P350.2 Write down without derivation or algebra the Shockley equations when both the donor and acceptor dopant impurities are partially ionized and mobile.

P350.3 Write down by inspection without derivation the Shockley equations if protons (positive hydrogen ion) and hydrogen atoms are present and mobile, and the protons can be trapped by the dopant acceptor and donor impurities.

P350.4 Show by inspection without derivation or algebra that the sixth Shockley equation (350.6), which I have called the (trapping) kinetic equation since formulating it in the 1960's to analyze 1/f and $1/f^2$ generation noise spectra in field-effect transistors and capacitance-current transient spectroscopy in p/n junctions, is really a continuity equation of the trap in disguise.

P3611.1 Draw the final-state interband thermal transition energy band diagrams of a direct material. How do they compare in clarity and utility with the initial-state diagrams shown in Fig.3611.1?

P3611.2 Draw the (i) initial-state and (ii) final-state transition energy band diagrams of a direct material.

P3612.1 Draw the initial-state E-x and E-k transition energy band diagrams and optical transition arrows for the optical generation (or optical absorption) transitions that correspond to the inverse of the radiation recombination transitions shown in Fig.3612.1.

P3612.2 Ge has an energy gap of 0.66eV and it is indirect, i.e., its conduction and valence band edges are not at the same k point. Demonstrate quantitatively that Ge is not a good light emitter and absorber because the change in electron momentum or k during an interband optical transition cannot be conserved by the participating photon. Lattice constant of Ge is $a=5.64613\times10^{-8}$cm.

P3613.1 Draw the final-state interband Auger-impact transition energy band diagrams of Fig.3613.1

P3622.1 Draw the final-state and initial-state transition energy band diagrams of the band-trap optical grt processes.

P371.1 A thin Si crystal is exposed to a pulse of light with $h\nu>E_G$. What is the time constant of the conductivity variation with time if only interband thermal and optical mechanisms are important.

P372.1 Derive the expression for the SRH hole lifetime in n-Si, (372.4).

P372.2 What is the total lifetime of electrons in a p-Si containing two trapping species?

P372.3 Give the algebra that leads to the minority carrier lifetime due to the interband optical mechanism in an n-type semiconductor. Invoke detailed balance at optical equilibrium, i.e. no external light. Then, expose the sample to a pulse of monochromatic light.

P381.1 An n-type Si crystal (10^{17} donor/cm^3) contains 10^{15} Au/cm^3. What are the low level and high level lifetimes of electrons and holes in this sample?

P381.2 Use the analysis procedure and physical model of the thermal and capture emission rates in section 381, obtain the expressions and discuss the physics of the radiative capture and photo emission rates at an optical trap.

P382.1 The indium acceptor in Si, $(E_T-E_V)=270$mV, was used as 4-micron infrared detector in satellite and space-craft applications since there is a atmospheric absorption window around this wavelength. The absorption cross section is about a Bohr radius of trapped hole, 10^{-16}cm^2. The solubility of indium in silicon crystal is about 10^{16}cm^{-3}. The infrared photon flux is 10^{14}cm^{-2}s^{-1}. How long would it take to detect 100 photons using an indium detector of 1 cm^2 area?

P383.1 Si, GaAs and CdS crystals are exposed to a yellow light of wavelength 5500A and intensity 10^{12} photons/cm^2-sec. Verify that the photon energy is 2.25eV. How many electron-hole pairs are generated per unit volume per second in each crystal? Assume that crystal is very thin and the light is absorbed uniformly in the very thin slice. Why do we need to assume the crystal to be very thin? (Use $\lambda=1.24/E$ and Fig.383.1.) (Partial answer: Si, 7×10^{15}cm^{-3}s^{-1}.)

P384.1 How many electron-hole pairs are generated per second in Si by interband impact generation at x_1 if $E(x_1)=3\times10^5$V/cm. If $N(x_1)=n_i=1.0\times10^{10}$cm^{-3}, what is the number of electron-hole pairs generated per unit volume per second? How does this compare with the optical generation rate obtained in problem 383.1? (Use Fig.384.1).

P384.2 Explain the very large generation rate obtained in the first part of problem 384.1. How many electron-hole pairs are created by an electron passing through a layer of 1μm thick at the scattering limited velocity?

P385.1 Electrons are slowly injected by tunneling into the SiO$_2$ from Si in a MOS transistor when a high electric field appears across the oxide. A small but significant fraction, about 10^{-4}, of the injected electrons will be captured by oxide electron traps in the SiO$_2$. The transistor will age and the MOS integrated circuit will fail when -10^{11}q/cm^2 charges are trapped in the oxide. Calculate the time to failure when 10V is applied to a 100A VLSI oxide. What is the maximum applied voltage to the oxide film for 10-year life?

P386.1 The electron binding energy at an oxide trap is found to be 1eV and the potential barrier height between the Si and the SiO$_2$ film grown on Si is 3.1eV. Electrons are slowly injected by tunneling from Si into a trap in the gate oxide of a MOS transistor when +1V is applied to a 100A-thick oxide film on the Si substrate. Suppose that the integrated circuit using this MOS transistor will fail when 10^{11}electron/cm^2 are trapped at the oxide electron trap. How long does it take to fail if there are 10^{12} cm^{-2} unoccupied traps initially? What is the maximum applied voltage to the oxide for 10-year life?

P386.2 Reconcile the two problems, P385.1 and P386.1, by estimating a oxide trap concentration for P385.1 and a trapping efficiency for P386.1.

Chapter 4
METAL-OXIDE-SEMICONDUCTOR CAPACITOR (MOSC)

Begin with the simplest physics
and
the largest application volume.

400 INTRODUCTION 312
  401 Fabrication of an Silicon VLSI MOSC, 314
  402 Ideal C-V Curves, 320
  403 Real C-V Curves, 323
410 CHARGE CONTROL MODEL OF MOSC 325
  411 Charge Control C-V Theory without Energy Band Diagram, 336
    - Depletion Capacitance, 338
    - High-Frequency Capacitance, 339
    - Low-Frequency Capacitance, 342
    - Accumulation Capacitance, 344
    - Flat-Band Capacitance, 344
    - Summary, 346
  412 Advanced Charge-Control C-V Theory, 347
    - Relating Electric Field to Potential at Semiconductor Surface, 347
    - Relating Surface Potential to Gate Voltage, 349
    - The Exact Low-Frequency MOS Capacitance, 353
    - Depletion and High-Frequency Capacitances, 353
  413 Energy Band Diagram of MOSC, 354
420 TRANSIENTS IN MOSC 363
  421 Capacitance Transients, 365
  422 Current Transients, 369
430 EXACT SMALL-SIGNAL EQUIVALENT CIRCUIT OF MOSC 374
499 BIBLIOGRAPHY AND PROBLEMS 375

400 INTRODUCTION

Figures 400.1(a) and (b) show the top and cross sectional view of a silicon metal-oxide-semiconductor capacitor (**MOS capacitor** or MOSC). It contains three heterojunctions between four layers, the metal gate (G node), oxide (SiO_2), **semiconductor body** or **substrate** (Si), and back metal (X node) for external low-resistant contact to the Si substrate or Si body. The X node label follows the usage in MOS transistors to be discussed in chapter 6. The MOS capacitor is one of the most important device structure in silicon transistors and integrated circuits and yet the simplest to understand among all solid state devices. Thus, it will be the first semiconductor device presented to the beginner in this introductory textbook. Its simple operation principle is because it passes no d.c. current (oxide is an ideal zero conductance insulator), therefore, both the diffusion and drift currents are zero and only electrostatics is needed. Its capacitance variation with applied d.c. voltage can be derived using the differential capacitance of the charge control method, $C=dQ/dV$ or $C=(dQ/dt)/(dV/dt)$ in which the charge density, Q, is derived from electrostatics.

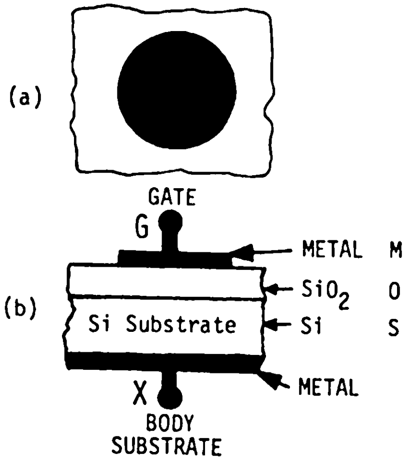


Fig.400.1 Physical picture of a MOS capacitor on silicon. (a) the top view of the silicon chip, and (b) the cross-sectional view.

Modern applications, such as the charge storage capacitor for one bit of information in DRAM (dynamic random access memory), have used MOS capacitors with two different dielectric layers to increase the capacitance and a polycrystalline Si gate buffer layer to prevent Al from reacting with and penetrating into the thin (< 100A) oxide layer. Such a capacitor would have a structure $Al/poly\text{-}Si/Si_3N_4/SiO_2/p\text{-}Si/Al$ (M/poly/N/O/S/M or MNOS as a short acronym).

The original MOSC structure was proposed by John L. Moll in 1959 [400.1] as a voltage-control capacitance, known as **varicap**. It was composed of an aluminum metal electrode on a thermally oxidized semiconducting silicon crystal like Fig.400.1. At that time, a diffused silicon p/n junction diode was used as the varicap to replace the mechanical variable capacitor in the frequency-selection tuned-LC circuit at the front end of AM-FM and CB radio, TV, garage opener and other receivers. Metal-oxide-silicon capacitor was suggested as a replacement since it passes no d.c. current due to the oxide insulator. Moll coined the name, **MOS Capacitor**, abbreviated by this author as MOSC to be consistent with 'MOST' for the MOS field-effect transistor.

Moll also suggested the use of the MOSC to monitor the quality of the oxidized silicon surface during fabrication. Moll's graduate student at Stanford, Lewis M. Terman, undertook a detailed doctoral research study on the properties of the bound states at the oxide/silicon interface of MOSCs which was reported in 1962 [400.2]. The bound states at a boundary or in the interfacial layer between two materials is known as the **interface traps**. Terman measured the frequency dependence of the small-signal capacitance as a function of the d.c. voltage applied to the MOSC. By comparing the experimental capacitance-voltage (C-V or MOSCV) data with theory, Terman obtained the density and the electron-trapping and hole-trapping time constants of the interface traps of thermally oxidized silicon. He further showed that the high-frequency C-V (HFCV) data gives the total density of the interface traps. Since the 1970's, the HFCV method has been universally used as the diagnostic method to monitor the fabrication processes of silicon VLSI (very large scale integrated) transistors and circuits because of its ease of use and lack of ambiguity. It is known as the **Terman method** or **Terman-Moll method**.

The first detailed demonstration of using the MOSCV data to study transistor grade oxides on oxidized silicon surface was presented by Grove, Deal, Snow and Sah in 1965 [400.3]. They were also the first to present the correct physics and theory of the HFCV curve [400.4]. With a 100KHz Tektronix model 130 LC meter modified by Sah to give an analog output proportional to the capacitance in order to trace the CV curve on a x-y recorder, they were able to monitor a large number of HFCV curves in many sodium contaminated MOSC's in 1965. From these data, they demonstrated that electrical instability in MOSC at room temperature was caused by the migration of positively charged sodium ions from the aluminum gate electrode into the oxide during the application of a high positive d.c. voltage to the aluminum gate of MOSC [400.5].

MOSC finds two additional device applications which dominate electronics today. Its most overwhelming application is as the **gate** or input electrode of the MOS field-effect transistor (MOST) [400.6] to be described in chapter 6. Thus, we shall use the word 'gate' frequently, such as the **gate capacitance** C_G or C_g, to denote the MOSC parameter, **gated oxide** to denote the oxide area covered by a metal or conductor gate electrode, and **gated oxidized silicon** to denoted the silicon surface area under the oxide/silicon interface and under the gate electrode. MOS capacitor is also the basic charge storage element of the charge-coupled device

(CCD) invented by Boyle and Smith at Bell Laboratories in 1970 [400.7]. CCD is composed of a closely spaced array of MOSC electrodes and its principal volume applications have been the imaging element in video camera, and the signal delay line in digital and analog circuits.

Intermediate and advanced topics on MOS device physics and technology are discussed in books [400.8-400.10] published during the 1980's. To keep abreast of the rapid new developments in MOS devices and integrated circuits in the 1990's, the Proceedings of the International Electron Device Meeting (IEDM) and International Solid-State Circuits Conference (ISSCC) are to be consulted.

[400.1] John L. Moll, "Variable capacitance with large capacitance change," IRE 1959 WESCON (Western Convention and Show) Record, part 3, pp.32-36, IEEE Service Center, 455 Hoes Lane, P.O.Box 1331, Piscataway, NJ.

[400.2] Lewis M. Terman, "An investigation of surface states at a silicon silicon dioxide interface employing metal-oxide-silicon diodes," Solid-State Electronics, 5(5), pp.285-299, Sept.-Oct., 1962.

[400.3] Andrew S. Grove, B. E. Deal, E. H. Snow, and C. T. Sah, "Investigation of thermally oxidized silicon surfaces using metal-oxide-semiconductor structures," Solid-State Electronics, 8(2), pp.145-163, February 1965.

[400.4] Andrew S. Grove, E. H. Snow, B. E. Deal and C. T. Sah, "Simple physical model for space-charge capacitance of metal-oxide-semiconductor structures," J. Applied Physics 33(8), pp.2458-2460, August 1964.

[400.5] Edward H. Snow, A. S. Grove, B. E. Deal, and C. T. Sah, "Ion transport phenomena in insulating films," J. Applied Physics, 36(5), pp.1664-1673, May 1965.

[400.6] C. T. Sah, "Characteristics of the MOS Transistor," IEEE Trans. ED-11(7), pp.324-345, July 1964.

[400.7] W. S. Boyle and G. E. Smith, "Charge coupled semiconductor devices," Bell System Tech. J. 49, pp.587-597, May 1970.

INTERMEDIATE TEXTBOOKS AND ADVANCED REFERENCES

[400.8] S. M. Sze, **Physics of Semiconductor Devices**, 2nd ed., chapter 7, "MIS diode and charge-coupled devices," pp.362-430, John Wiley & Sons, Inc., 1981.

[400.9] L. E. Katz, "Oxidation," chapter 4 in **VLSI Technology**, edited by S. M. Sze, 2nd ed., McGraw-Hill Book Company, 1988.

[400.10] I. W. Boyd, et.al., "Oxidation," chapter 16; C. T. Sah, et.al., "Insulating layers on Si substrate," chapter 17; and H. R. Philipp, "Silicon Oxides: Optical Functions," chapter 28; all in **Properties of Silicon**, INSPECT, The Institution of Electrical Engineers, London, 1988. 445 Hoes Lane, P.O.Box 1331, Piscataway, NJ 08855-1331.

---

401 Fabrication of an Silicon VLSI MOSC

The twelve fabrication steps of an aluminum/silicon-dioxide/silicon MOS capacitor using the latest silicon integrated circuit manufacturing technology are shown in Fig.401.1(1) to Fig.401.1(12). These steps are described briefly. The figures show the device structure after each fabrication step has been completed.

Figures (1) to (6) show only the cross sectional views. Figures (7) to (12) shows both the cross sectional views and the top views.

Only one capacitor is shown from an Si wafer of 3 to 8 inch diameter. In research, the metal electrode (dark dot) capacitor has a diameter of about 7.5 mil (1 mil = one thousandth of an inch = 25.4 micrometers) and MOSC chip is a 0.1-inch by 0.1-inch square with an area of $A_{chip}=0.1^2=0.01\text{-inch}^2=10\text{x}10\text{mil}^2$. Thus, on a 3-inch diameter Si wafer, there are approximately $\pi d^2/4A_{chip}$ = $\pi\text{x}3^2/4(0.01)^2$ = 70,000 chips or capacitors. On a 8-inch diameter wafer, there are about 502,000 0.1x0.1 inch square chips or half-of-a-million MOS capacitors.

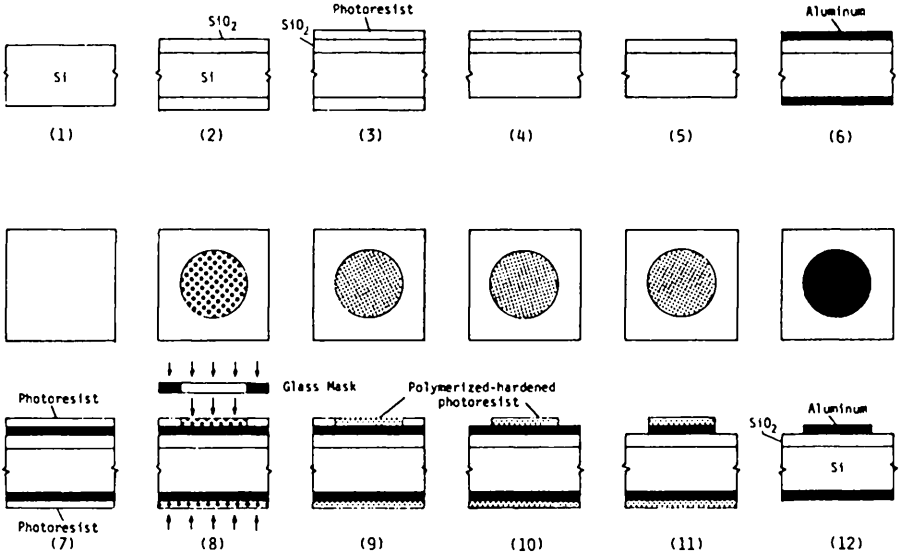


Fig.401.1 The twelve fabrication steps of an Al/SiO$_2$/Si MOS capacitor using the latest silicon integrated circuit manufacturing technology. The steps are described in the text.

The maximum capacitance of a MOSC is the oxide capacitance, $C_O=\varepsilon_o A_g/x_o$ where ε_o is the permittivity of the oxide (for SiO$_2$, $\varepsilon_o=3.9\text{x}8.854\text{x}10^{-14}$ F/cm), x_o is the thickness of the oxide, and A_g is the area of the metal electrode, known as the gate electrode. In research, the MOSC is designed to have a maximum or oxide capacitance value of 10, 30 or 100 pF in order to utilize the full scale of a capacitance meter for maximum accuracy and sensitivity. Smaller values are desirable to study areal inhomogeneity effects but then stray capacitances may become significant compared with the MOS capacitance and would introduce significant measurement errors. For example, the oxide capacitance of the Si MOSC just illustrated would be 10.0 pF if the oxide thickness is 1000A (0.1 μm or 100nm) and if the metal gate area has a diameter of 192 μm (or 7.56mil), which gives $C_o = \varepsilon_o A_g / x_o = 3.9 \mathrm{x} 8.854 \mathrm{x} 10^{-14} \cdot \pi \cdot (192 \mathrm{x} 10^{-4}/2)^2 / 1 \mathrm{x} 10^{-5} = 10.0\mathrm{pF}$.

MOSC's used to monitor the manufacturing steps of a Si VLSI circuit are usually placed at several test-pattern locations on a 6- or 8-inch Si wafer. Its oxide thickness is identical to that used in the gate of a MOS transistor, ranging from 1000A for large power MOS transistors to 100A for the latest submicron Si VLSI integrated circuits. The area of the MOSC can be scaled in accordance with the oxide thickness so that the oxide capacitance is about 10 pF to allow accurate capacitance measurement.

Figure (1) of Fig.401.1 shows the cross sectional view of a 500-micron thick Si wafer. Its two surfaces are polished by a chemical etchant to remove the mechanical damage and impurities left on the saw-cut Si wafer. The top surface is chemically polished to a mirror finish with surface flatness of about 1/4 of a wavelength of light, or about 0.1 micron, in order to met the submicron tolerance requirement of VLSI circuits. Still better flatness is required when near-field optical (100A resolution) and x-ray lithographies are used to fabricate and manufacture in large volume sub-1000A transistors and integrated circuits in the next century. The bottom surface is chemically polished only to a dull finish.

Figure (2) shows a 1000A oxide film (SiO_2) is formed on both the upper and lower surfaces of the Si wafer. The oxide film is grown on the Si surfaces by thermal oxidation of the Si wafer inside a furnace tube at about 1000°C for two hours with oxygen flowing through the furnace tube. The furnace tube is made of fused silica. The cleanness of the furnace tube and purity of the oxygen gas are key factors which determine the electrical quality of the oxide film. The thermal rate of growth of the SiO_2 film on the Si surface is limited by atomic diffusion of both the oxygen atoms from the ambient through the already formed oxide film to the oxide/silicon interface and the outdiffusion of Si atoms to the surface of the already formed oxide. The oxygen at the SiO_2/Si interface then rapidly reacts with a Si atom to form another SiO_2 bond while the outdiffused Si atoms at the SiO_2 surface (or the O_2/SiO_2 interface) rapidly reacts with an oxygen atom to form another layer of oxide. Thus, during thermal oxidation of silicon, the oxide film grows into the silicon and also outwards on top of already formed oxide. The inward/outward growth ratio is about 4/6 and this is shown in figure (2) compared with figure (1). In particular, the upper SiO_2/Si boundary in figure (2) is below the original upper Si surface of figure (1).

The oxidation rate can be increased and the oxidation time shortened if the oxygen ambient contains a partial pressure of water (H_2O). Oxidation rates of Si in dry and wet oxygen and pure steam have been measured experimentally using optical interferometry to measure the oxide thickness. Figure 400.2 gives the oxidation rate, $x_0/\surd t_0$, as a function of the oxidation temperature for various partial pressures of oxygen (oxygen+argon) and water vapor (oxygen+water). t_0 is the oxidation time in hours. As an example of using the figure, a 1000A thick oxide can be obtained in 160 seconds at 1000C in steam (760mm of H_2O).

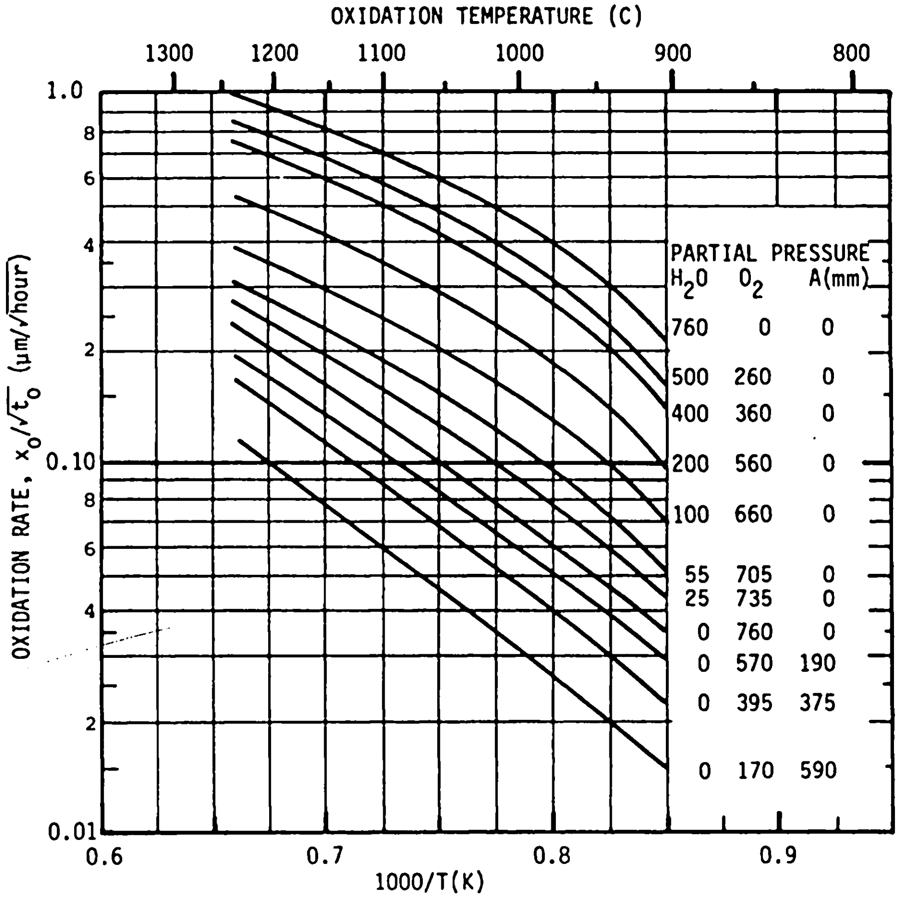


Fig.401.2 The oxidation rate of silicon in engineering unit, $x_0/\surd t_0$ ($\mu m/\surd$Hour), as a function of oxidation temperature with oxygen partial pressure in argon (lower curves) and water partial pressure in oxygen (upper curves) as the parameter.

The top oxidized surface is then coated with a layer of photoresist film to prevent the top oxide layer from being etched as indicated in Fig.401.1(3). The bottom oxide film is then etched off chemically as indicated in Fig.401.1(4). The photoresist film on the top surface is then removed as indicated in Fig.401.1(5).

Highly pure aluminum metal (99.9999% or better purity) is then evaporated inside a high vacuum belljar (less than 10^{-6} mm-Hg) onto both the oxidized top Si surface and the bare bottom Si surface as shown in Fig.401.1(6). The aluminum metal is shown as darkened area in these figures. The next step is to coat the metallized or aluminized top and bottom surfaces with a layer of photoresist in order to define the geometry of the metal gate electrode (a circular dot in this example) on the top surface. The end result is shown in Fig.401.1(7).

To obtain the desired gate geometry pattern, the top photoresist film is exposed to light through a glass mask which has the desired pattern (circular dots) on the mask as indicated in Fig.401.1(8). The bottom photoresist is uniformly exposed to light without a glass mask. The photoresist is then heated (baked in an oven) at about 150C-200C as indicated in Fig.401.1(9). The heat polymerizes and hardens the photoresist film. The exposed area of the photoresist film will also adhere to the Al metal while the unexposed area on the photoresist on the top surface are etched off chemically, resulting in the pattern shown in Fig.401.1(10). The wafer is then immersed in an aluminum etchant to remove the exposed aluminum, resulting in Fig.401.1(11). The photoresist prevents the photoresist-covered aluminum on both the top and bottom surfaces from being etched away. A highly pure phosphorus acid (HPO_3), known as MOS grade which is purer than the electronic grade, is used as the etchant today. During the infancy of the Si VLSI technology, around 1956-1959, NaOH was used to remove the aluminum with disastrous results - the capacitance-voltage curves of the MOSC's thus made were extremely unstable. It was quickly determined that Na ions from NaOH are incorporated into the oxide and the Na ions are extremely mobile in SiO_2 even at room temperatures [400.5].

The patterned photoresist film on the top surface and the blanket photoresist film on the bottom surface are then removed by a solvent which can dissolve the polymerized and hardened photoresist films. The Al electrode on the top surface and the Al contact on the bottom surface are then exposed as indicated in Fig.401.1(12), giving about 70,000 aluminum dots or MOSC's on the surface of a 3-inch or 500,000 on an 8-inch oxidized Si wafer.

At this stage, additional processing steps are usually taken in MOS research. However, in manufacturing to monitor a VLSI production process or to develop a new VLSI technology, there are usually no additional processing steps. Capacitance-voltage measurements are made on each MOSC by pressing a metal probe to make electrical contact with the aluminum gate dot while the aluminized bottom surface sits on a metallic platform which serves as the other terminal. This probing method provides rapid feedback for production control but limits the measurement accuracy due to temperature fluctuation and electrical noise from the atomically poor pressure contact on the dot and back surface.

In basic research on the properties of the oxide and their effects on the performance and reliability of the MOS transistors and integrated circuits, accurate measurements are needed which require constant temperature and low electrical noise. These requirements can be attained by housing the MOSC inside a metal package such as a transistor header illustrated in Fig.401.3 whose metal can is removed. Thus, the 3-inch to 8-inch Si wafer with the aluminum MOSC's on the surface are diced into 0.1x0.1 inch^2 square chips or **dies** using a computer-controlled saw with a diamond cutting wheel. A chip or die is then soldered onto a two-lead (or two-terminal) gold-plated transistor header in a controlled ambient (usually highly pure and dry nitrogen gas) using an alloy whose melting point is between 70-400C, known as **die attach** or **chip bond**. One of the two terminals is in contact with the gold-plated transistor header, so this terminal is already connected to the aluminized back surface of the MOSC. The aluminum dot on the top surface is connected to the unattached terminal-post by a fine (1-mil diameter) gold wire known as **wire bond** or **wire attach**. One of the two metal-wire bonding methods is used to attach the gold wire to the aluminum dot and the unattached terminal-post of the transistor header. The two methods are the thermo-compression and ultrasonic bonding methods. The thermo-compression method requires a controlled ambient to prevent oxidation and reaction because the chip and header is heated up to 400C to help break through the metal oxide films between the two metals (Al_2O_3 in the case of Al-Au bond). The ultrasonic method uses high frequency mechanical vibration to break up the oxide film at room temperature, thus, it is the preferred bonding method because no heating is required. However, the pressure and vibration must be controlled so as not to damage the thin oxide.

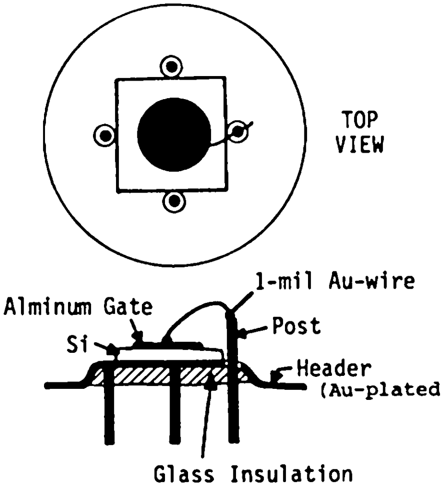


Fig.401.3 A Si MOS capacitor in a gold-plated transistor header with the can removed. It is used in accurate measurements of the fundamental electronic properties of the oxide, semiconductor, and their interface.

402 Ideal C-V Curves

Three typical capacitance variations with d.c. voltage can be observed on one MOS capacitor at three different measurement conditions. The three MOS **capacitance-voltage** (C-V or CV) curves are shown in Fig.402.1(a) on an n-Si substrate, and in Fig.402.1(b) on a p-Si substrate. The three curves are the **low-frequency capacitance-voltage** (LFCV) or differential capacitance-voltage characteristics, C_{LF}-V_G or C_{lf}-V_G; the **high-frequency capacitance-voltage** (HFCV) characteristics, C_{HF}-V_G or C_{hf}-V_G; and the **depletion capacitance-voltage** characteristics (DCV), C_D-V_G or C_d-V_G. The circuits used to measure these MOSCV curves are also shown which indicate the polarity of the **applied d.c. bias voltage**, V_{DC}, and the two terminals at which the device capacitance, C, is measured by a capacitance bridge or capacitance meter.

At high positive gate biases on the n-Si MOSC and high negative gate biases on the p-Si MOSC, all three CV curves approach a constant value, C_O. One can easily verify by area and thickness measurements that the observed C_O is the capacitance of a parallel-plate capacitor having the oxide layer as the dielectric film. C_O is known as the **oxide capacitance**. The asymptotic constant capacitance behavior can be easily understood from simple electrostatics. Consider the MOSC on n-Si shown in Fig.402.1(a). At large positive d.c. gate bias, many electrons are attracted to the semiconductor surface under the aluminum gate electrode from the interior or bulk region of the n-Si. The accumulation of a high concentration of electrons at Si surface makes the Si surface metallic-like and acting as the second conductor electrode of the oxide capacitor opposite to the first or aluminum gate electrode. This gives the constant capacitance C_O. It is known as the majority carrier accumulation range of the applied d.c. voltage, or simply the **accumulation range** or **surface accumulation**. It is analogous to applying a **forward** d.c. **bias** voltage to a rectifying diode such as a positive d.c. voltage to the p-type terminal of a p/n junction diode, or a positive d.c. voltage to the metal terminal of a metal/n-semiconductor junction diode. The difference is that MOSC passes no d.c. current unless the oxide is porous or so thin that electrons or holes can tunnel through the thin oxide layer. Thus, on n-Si MOSC, **positive** gate bias will be called the **forward bias**, and **negative** gate bias, the **reverse bias**. Similarly, on p-Si MOSC, negative gate bias is forward bias while positive gate bias is reverse bias.

Under **reverse bias**, the MOSCV characteristics have three different voltage dependences for the following reason. Consider the p-Si MOSC. The reverse (positive) gate bias repels the majority carriers (holes) from the gated Si surface and simultaneously attracts the minority carriers (electrons) to the gated Si surface. This is known as **surface inversion** because there are more minority carriers at the Si surface than majority carriers, causing the surface conductivity to invert to the type opposite to that of the bulk. The key factor that controls the three different MOSCV behaviors is the **rate of supply** of the minority carriers (via generation and

diffusion) to the surface inversion layer compared with the rate of change of the signal voltage.

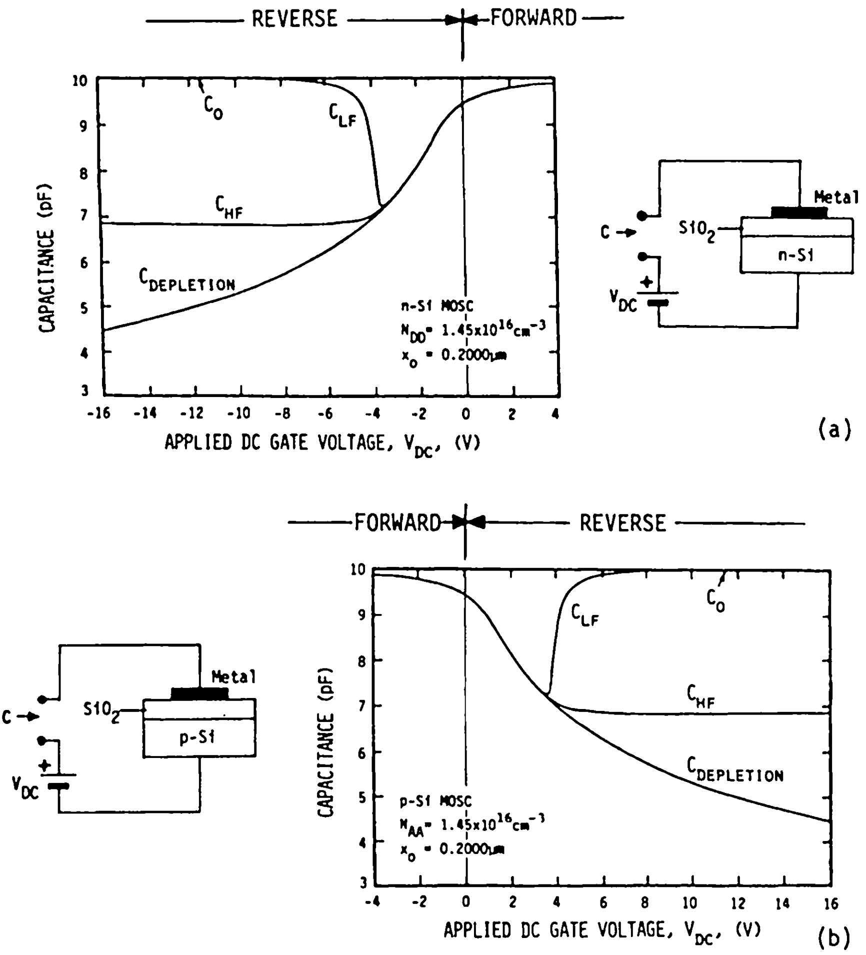


Fig.402.1 The MOS CV curves of ideal MOS capacitors on (a) n-Si and (b) p-Si.

The two higher CV curves, LFCV and HFCV, are observed when the d.c. gate current is zero. Thus, they are **equilibrium** (thermal and electrical equilibrium) curves. They are commonly observed in Si MOSC's since the SiO_2 used as the gate oxide is usually a nearly perfect insulator. This leakless SiO_2

permits the accumulation of a high steady-state (equilibrium) concentration of minority carriers (electrons in p-Si MOSC) at the gated Si surface of the SiO_2/Si interface so that the Si surface layer is inverted (to n-type on a p-Si MOSC) by applying a d.c. gate bias of the proper polarity (positive on p-MOSC).

The HFCV curve is observed in the dark at 1 MHz or a sufficiently high signal frequency when two conditions are met: (i) the MOS oxide is a good insulator so that a high d.c. steady-state concentration of minority carriers can accumulate at the oxide/Si interface, and (ii) minority carriers cannot be supplied to and extracted from the oxide/silicon interface rapidly enough (by diffusion or generation-recombination) to respond to the signal voltage variation at 1 MHz.

The LFCV curve is observed when the minority carriers can be supplied rapidly or the signal frequency is sufficiently low. The rise of the capacitance towards C_O in the inversion range (positive V_G on p-MOSC and negative V_G on n-MOSC) comes about because the minority carrier density in the inversion layer can now be changed rapidly enough to follow the signal frequency. There are several high-speed source and sink of minority carriers: (i) optical generation of minority carriers, (ii) existence of a built-in semiconductor region of the minority carrier conductivity type which is partially covered by the gated oxide, such as the diffused source or drain junction of a MOS transistor. (i) does not give a strictly equilibrium LFCV curve since there is an excess of electrons and holes from optical generation or $NP > n_i^2$, but if the light intensity is sufficiently low, then $NP \approx n_i^2$ and a nearly equilibrium LFCV curve can be observed. (ii) gives a strictly equilibrium LFCV only when the minority carrier signal delay along the inverted surface layer is small compared with the reciprocal frequency and the minority-carrier source (source and drain junction of a MOST) is electrically connected to the majority-carrier substrate or bulk.

The depletion capacitance or DCV curve is observed if the minority carriers cannot be accumulated at the oxide/silicon interface or are generated so slowly that few are generated and accumulated at the gated Si surface before the C-V trace is completed. There are three possibilities: (i) the oxide or insulator is leaky and the minority carriers are leaking off to the gate faster than they can be generated; (ii) the minority carrier generation rate is so small that few minority carriers are generated before the CV trace is completed such as a measurement made at low temperatures and the MOSC is shielded from light, (iii) there is a built-in minority-carrier semiconductor drain region under the gated oxide which is biased in a polarity that drains off the minority carriers, such as the diffused drain of the MOST. (i) will give a nearly theoretical DCV curve, however, highly conductive insulator films usually conduct bidirectionally so that the majority carrier may not accumulate as much in the positive gate voltage range resulting in a lower accumulation capacitance than C_O, and furthermore high conductivities in thin insulator films are often due to pin holes which would distort the DCV curve significantly.

403 Real C-V Curves

Experimental CV curves are different from the ideal theory of Figs.402.1(a) and (b). The experimental CV curves (dark curves) are shown in Fig.403.1(a) to (d) for n-MOSC and in Fig.403.2(a) to (d) for p-MOSC. They can be compared with the ideal theory (light curves).

A **parallel shift** of the CV curves along the d.c. applied voltage axis from the ideal theoretical CV curves, shown in Figs.403.1(a) and 403.2(a), indicates the presence of some trapped space charges which are distributed in the oxide layer and are time-invariant during the measurements of the CV curves. These space charges in the oxide are known as the **oxide charges** or **charged oxide traps**. They are frequently called 'fixed oxide charges' by practicing engineers, a misnomer, since they are nearly constant during measurement. However, they are neither fixed nor constant in space. Their migration is the vary cause of electrical instability in Si MOS transistors and integrated circuits. Their time dependence, arising from controlled injection-then-capture and emission-then-extraction of electrons and holes at the traps in the oxide, is used to write, store and erase a bit of information in some proposed Si MOS memory cells.

Figure 403.1(a) shows the **negative gate voltage shift** of the three MOSCV curves on n-Si by a **positive oxide charge** (positive space charge) or when the charges distributed in the oxide layer gives a **net positive** space charge seen from the Si. Figure 403.1(b) shows the **positive gate voltage shift** of the three CV curves when there is a **negative oxide charge** or a **net negative** space charge in the oxide. The corresponding curves for MOSC on p-Si are shown in Figs.403.2(a) and (b). Notice that the voltage shift direction is the same in p-type and n-type MOSC. It depends only on the sign of oxide charge and not on the sign of the charge of the donor and acceptor dopant impurities.

Distortion of the CV curves along the voltage axis compared with ideal theoretical CV curves are shown in Fig.403.1(c) for n-Si and Fig.403.2(c) for p-Si MOSCs. It is an indication that some charged traps are present at the oxide/Si interface or in the Si surface space-charge layer. These are known as **charged interface traps**, interface trapped charges, charged surface states, and surface state charges. We shall use the term **interface traps**. The distortion is caused by the voltage dependences of the density of the trapped interface charge when the applied d.c. gate voltage is varied. This very distortion or voltage dependence is used to measure the interface trap density experimentally known as the Terman method.

In general, voltage shift and distortion appear together in one CV curve, indicating the presence of both charged oxide and interface traps. Terman [440.2] separated the oxide and interface traps by measuring the CV curves at many frequencies. New measurement techniques have been developed recently to separate the oxide and interface traps by varying the oxide and interface densities individually via electron or hole injection from the Si substrate into the oxide.

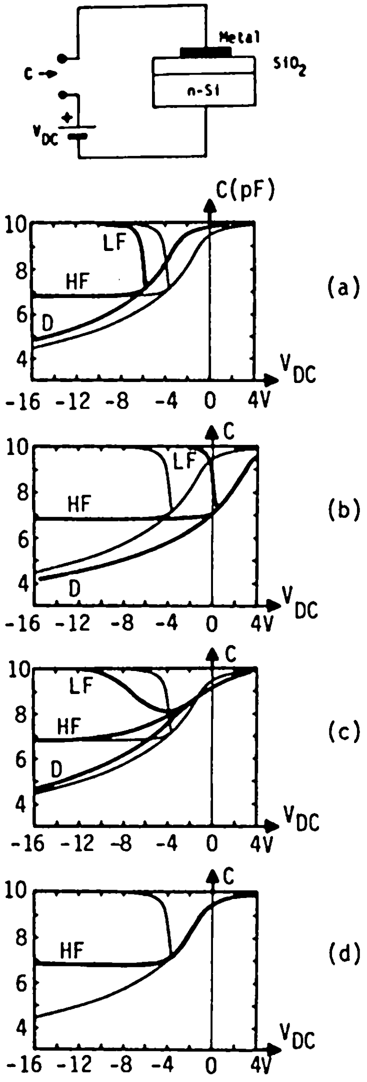


Fig.403.1 The CV curves on n-Si MOSC
(a) With positive oxide charge.
(b) With negative oxide charge.
(c) With charged interface traps.

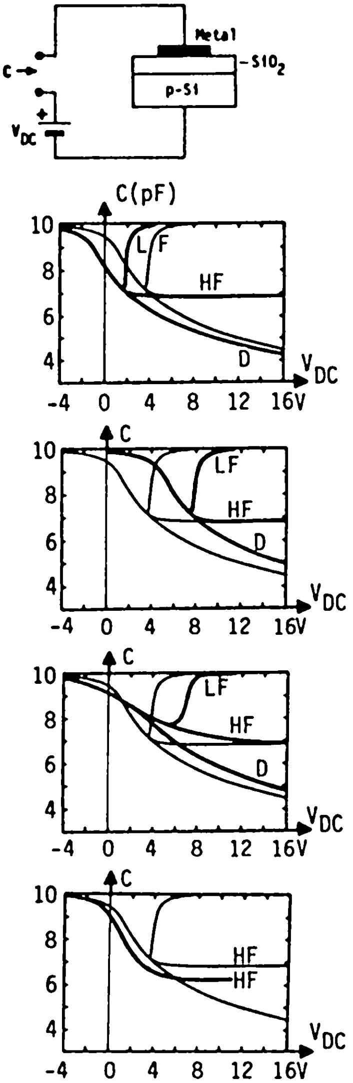


Fig.403.2 The CV curves on n-Si MOSC
(a) With positive oxide charge.
(b) With negative oxide charge.
(c) With charged interface traps.
(d) With B acceptor deactivated by hydrogen.

Finally, the low capacitance plateau of the HFCV curve can be shifted downwards in p-Si as indicated in Fig.403.2(d) when it is exposed to atomic hydrogen. However, a corresponding upward shift on n-Si MOSC, predicted by this author, has not been unambiguously observed as indicated in Fig.403.1(d). The downward shift in p-Si MOSC's is due to the deactivation of the boron acceptor by the atomic hydrogen forming the Si•H•B bond. It was first discovered in 1972 and explained in 1982 [403.1, 403.2]. It is known as the **acceptor hydrogenation effect** and is one of the mechanisms that can cause serious operating instability of Si MOS transistors and failure of integrated circuits.

[403.1] C.T.Sah,Y.C.Sun and J.T.Tzou, "Deactivation of boron acceptor in silicon by hydrogen," Applied Physics Letters, 43, pp.204-106, July 15, 1983.

[403.2] C.T.Sah, S.C.Pan and C.H.Hsu, "Hydrogenation and annealing kinetics of group-III acceptors in oxidized silicon," Journal of Applied Physics, 57(12), pp.5148-5161, June 15, 1985.

---

410 CHARGE CONTROL MODEL OF MOSC

The capacitance-voltage characteristics of the MOSC can be analyzed by the charge control model, $C=dQ/dV$, without using the energy band diagrams. This is an old method employed by earlier authors in introductory textbooks. It relies only on the Poisson equation and the Boltzmann relationship between the carrier concentration and the electric potential. However, the E-x energy band diagram can be used to advantage, by supplying the underlying physics, guiding the mathematical derivation, and relating the C-V curves to the fundamental material parameters of the gate conductor, oxide, and semiconductor such as the electron affinity, work function, and Fermi energy position and dopant concentration.

The charge control result is deficient because it is based on one-lump which does not take into account the signal delay due to diffusion and drift in the MOSC device structure. This deficiency is inherent in the one-lump circuit model which ignores the signal propagation delay in a multi-lump or distributed transmission line. Furthermore, it cannot predict charge injection, migration, generation-recombination-trapping-tunneling in the oxide and semiconductor surface layers. Some of the shortcomings can be alleviated by the use of the E-x energy band diagram. Nevertheless, the one-lump charge-control model will be used in the following sections in order to focus on physics and minimize mathematics. Analyses based on the distributed model are given in journal articles and future advanced textbooks. A general charge control analysis is given in the remainder of this section. In the next section, 411, the MOSCV curves are derived without using the energy band diagrams. In section, 412, the E-x energy band diagram is used to relate the C-V curves to the material parameters. In section 413, a step-by-step construction of the E-x energy band diagram is described.

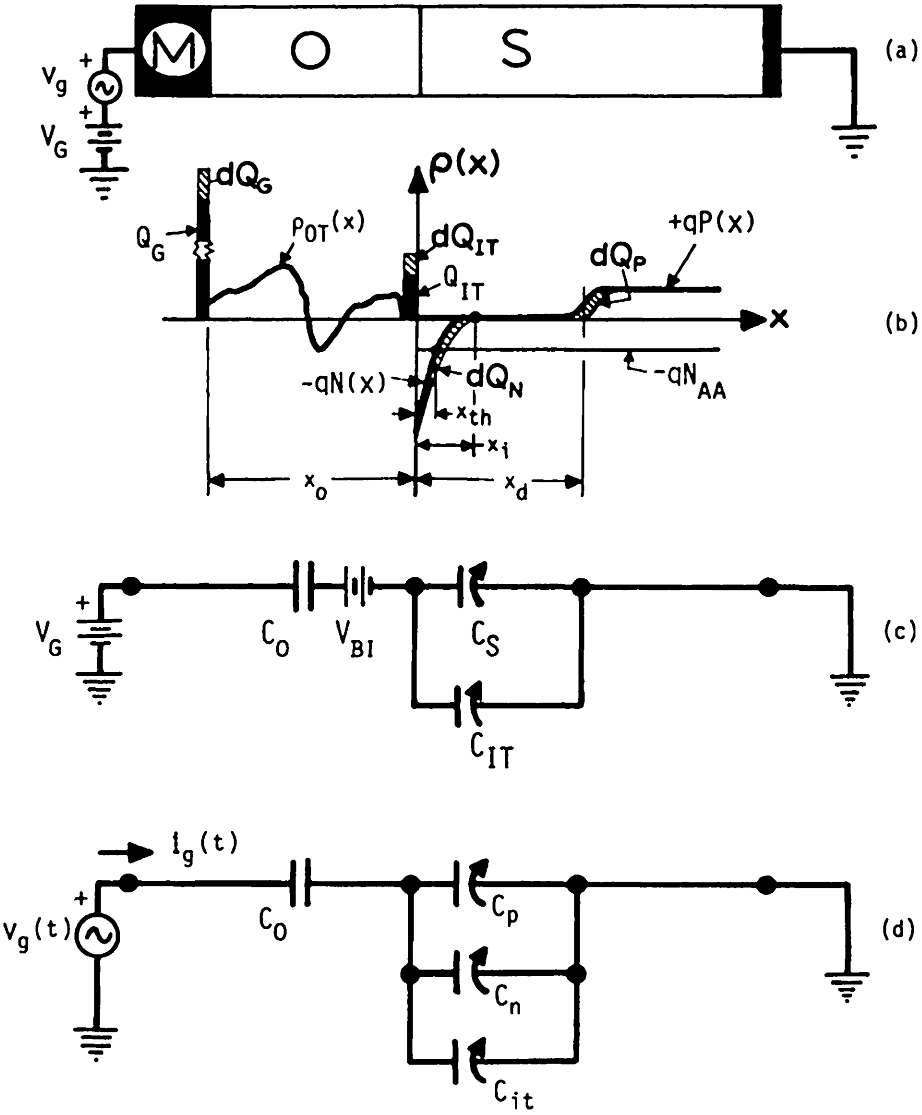

M
O
S
(a)
V_g
V_G
+
dQ_G
Q_G
$\rho(x)$
$\rho_{OT}(x)$
dQ_{IT}
Q_{IT}
$+qP(x)$
dQ_P
x
(b)
$-qN(x)$
dQ_N
$-qN_{AA}$
x_{th}
x_i
x_o
x_d
V_G
C_O
V_{BI}
C_S
C_{IT}
(c)
$I_g(t)$
$V_g(t)$
C_O
C_p
C_n
C_{it}
(d)

Fig.410.1 Schematic diagrams of a 1-dimensional MOSC model. (a) The cross sectional view where M = metal or conductor, O = oxide or insulator, S = silicon or semiconductor. (b) The space-charge density distribution. (c) The d.c. equivalent circuit model. (d) The charge-control small-signal equivalent circuit model.

To begin the charge control analysis, we decompose, via intuition, the MOS capacitor into two capacitances in series. These are the **oxide capacitance**, C_o, and the **semiconductor space-charge layer capacitance**, C_s. Both of these are capacitance per unit area since we are considering the one-dimensional model. The capacitance can be given in the unit (pF/cm^2) or in a more convenient unit for the problem on hand such as (fF/μm^2) of submicron Si VLSI circuits. The oxide capacitance is a constant, independent of the applied d.c. gate voltage. It is given by the simple parallel-plate formulae $C_o=\varepsilon_{ox}/x_o=K_o\varepsilon_0/x_o$ where K_o is the dielectric constant of the insulator, ε_0 is the permittivity of free space, and x_o is the thickness of the insulator. In order to simplify notation, one symbol and one subscript are used for the dielectric constant of the material: ε_o for the oxide in place of the double subscripted $\varepsilon_{ox}=K_o\varepsilon_0$; $\varepsilon_s=K_s\varepsilon_0$ for the semiconductor; and ε_0 (zero for subscript) for the permittivity of free space or vacuum.

The semiconductor capacitance, C_s, varies with the applied d.c. voltage, V_G. It is this variation that is responsible for the voltage dependence of the MOSCV curves. To obtain the voltage dependence of the semiconductor capacitance, we employ the charge control model to evaluate $C=dQ/dV$. Figure 410.1 shows (a) the cross-sectional view, (b) the space-charge density distribution, (c) the d.c. and (d) small-signal equivalent circuits. Notice that these figures are lined up vertically to illustrate the location of the d.c. steady-state charge distribution and the instantaneous movement and variation of the charges with time in a period comparable with the small-signal period. This vertical lineup is particularly helpful to visualize the charge control analysis of the MOSCV curves. We shall explain each of these figures first and then develop the engineering analyses of the MOSCV curves in the following paragraphs.

Figure 410.1(a) shows the cross-sectional view of a thin slice of the MOSC cut through the thickness of the Si wafer. The applied d.c. bias voltage and the a.c. small-signal test voltages are also shown.

Figure 410.1(b) shows the spatial variation of the space-charge density for a large positive applied gate voltage, $V_G>>0$. Q_G is the total areal charge density (C/cm^2) on the gate conductor. $\rho_{OT}(x)$ is the volume density (C/cm^3) of the space charge trapped in the oxide, i.e., the charge density of the oxide traps. Frequently, $\rho_{OT}(x)$ is represented by its moment with respect to the gate-conductor/oxide interface, Q_{OT}(C/cm^2)$=\int x\rho_{OT}(x)dx$ (integration over the thickness of the oxide $x=0$ to $x=x_o$). It is the total effective oxide charge seen by the semiconductor and known as the **oxide charge** or **oxide trapped charge**. Q_{IT} (C/cm^2) is the total areal charge density residing at the traps located at oxide/semiconductor interface. N(x) is the volume density or concentration of the minority carriers (electron/cm^3 in p-type MOSC) that are distributed in the semiconductor surface layer while P(x) is that of the majority carriers (hole/cm^3). N_{AA} is the concentration of the ionized (negatively charged) p-type acceptor dopant impurity and is assumed spatially constant. The **inversion layer**, $0<x<x_i$, shown in the figure is the semiconductor

surface layer where $N(x) > P(x)$ or the p-type semiconductor is inverted to n-type by the positive voltage applied to the gate. Its boundary location or thickness, x_I, is obtained from $N(x_I) = P(x_I) = n_i$ and $x_I \simeq (0.414/1.414)x_d$ where x_d is the thickness of the surface space-charge layer show in figure (b). A very thin strong inversion layer, $0 < x < x_{th}$, is also defined in the figure (thickness is much thinner than drawn). It is the layer where $N(x) \geq P_B \simeq N_{AA} >> P(x)$. Its thickness is denoted by x_{th}. The subscript stands for threshold (coined by this author in 1964, see chapter 6) since $x_{th} = 0$ is the threshold condition for a large current to begin to flow in the surface channel of the MOS field-effect transistor to be discussed in chapter 6. Since $V_G >> 0$ is assumed, holes (majority carriers) are pushed away from the oxide/silicon interface, resulting in a surface space-charge layer of thickness x_d which is depleted of majority carriers, known as the majority-carrier depletion layer. It contains a extremely high concentration of minority carriers near the oxide/semiconductor interface but frequently the word majority-carrier is omitted. The depletion layer boundary x_d is not sharply defined as the concise definition of x_I and x_{th} just given. Roughly, it is given by $P(x_d) \simeq P_B/2 \simeq N_{AA}/2$, but little error is introduced if $P_B/10$ or $9P_B/10$ is used because $P(x)$ drops abruptly from P_B to 0 inside the space-charge layer due to the exponential dependence (Boltzmann relationship) on $V(x)$, $P(x) = P_B \exp[qV(x)/kT]$, and because $V(x)$ changes by kT/q in a small distance in the space-charge layer.

Figure 410.1(b) also shows the **differential** change of the charges due to a change of the d.c. gate voltage from V_G to $V_G + dV_G$, i.e. $v_g(t) = dV_G u(t)$ where $u(t)$ is the unit step function. For example, a positive gate voltage step produces the following charges: an increase or positive dQ_G on the gate, an increase or positive dQ_{IT} in the interface states, a magnitude increase or negative dQ_N for electrons, and a decrease or negative dQ_P for holes.

Figure 410.1(c) is the d.c. equivalent circuit model of the MOSC which has not been given before. V_G is the d.c. voltage applied to the gate. V_{BI} is the built-in potential drop across the entire MOSC due to the presence of charged oxide and interface traps and a work function difference between the metal, oxide and semiconductor layers. The formulae of V_{BI} will be derived in a later section. C_o is the oxide capacitance per unit area. C_S is the d.c. capacitance of the semiconductor defined by $C_S = Q_S/V_S$ where Q_S is the areal density of the net space-charge integrated through the semiconductor and V_S is the total potential drop through the semiconductor. Q_S is a function of V_S since it depends on the spatial distribution of the electrons and holes or $P(x)$ and $N(x)$. C_{IT} is the d.c. capacitance of the charges stored in the interfacial traps at the oxide/semiconductor interface, defined by $C_{IT} = Q_{IT}/V_S$, where Q_{IT} is the areal density of the total charge residing in the interface traps. Q_{IT} is also a function of V_S since the amount of charges trapped at the interface traps depends on the electron and hole concentrations at the interface, $P(x=0) = P_S(V_S)$ and $N(x=0) = N_S(V_S)$. The d.c. model gives charge conservation or the electrical neutrality relationship, i.e., the Gauss law applies to the semiconductor between its two metal contacts. It is

$$0 = Q_G + Q_{OT} + Q_{IT} + Q_S = C_oV_O + Q_{OT} + C_{IT}V_S + C_SV_S.$$

Q_{OT} is the total fixed charge density in the oxide, known as oxide trapped charge. It is part of the built-in potential, represented by the battery symbol and labeled V_{BI} in figure (b). Its other part comes from the interface charge, Q_{IT}, and the work function difference between the metal, oxide, and semiconductor layers.

Figure 410.1(d) is the small-signal equivalent circuit of the MOSC from the charge-control model. It also shows the signal generator, $v_g(t)$, connected to the gate and the direction of the small-signal gate current, $i_g(t)$. C_o is the same oxide capacitance per unit area as in the d.c. model given in Fig.410.1(c). C_p is the small-signal charge-storage or differential capacitance of holes stored in the semiconductor surface space-charge layer, defined by $C_p = -dQ_P/dV_S$ where the negative sign comes from our choice of coordinates of measuring V_S. C_n is the differential capacitance of electrons, $C_n = -dQ_N/dV_S$. C_{it} is the differential capacitance of the charges stored in the interface traps at the oxide/semiconductor interface, given by $C_{it} = -dQ_{IT}/dV_S$.

The MOS capacitance, C_g, is a small-signal circuit quantity. It represents the response of the stored (distributed) mobile charges (electrons and holes) in the device when a small-signal voltage, current, other excitation, disturbance, perturbation, or force (heat, mechanical stress, light or particle radiation) is applied to the device. They can be applied to the two contact leads directly (current or voltage) or to the device's physical body via exposure to electromagnetic radiation (photon or light) or surface-ambient atomic-molecular collision (heat).

The term **response** comes from the small-signal time-dependent change of the current or voltage that is responding to the applied excitation or force

The term **small-signal** means that the excitation or force is sufficiently small so the relationship between response and force is **linear**. Semiconductor devices are highly nonlinear because the concentration of the charge carriers or mobile charges (electrons and holes) varies tremendously with a small variation of the electric potential. The variation is **exponential** as indicated by the Boltzmann factor, $\exp(\pm qV/kT)$, discussed in chapter 2. The linearity condition of the small-signal model requires that the maximum change of the electric potential, δV is small compared with the thermal voltage, kT/q, i.e., $\delta V << kT/q$ or $\delta V << 25$ millivolts at room temperature (~ 300K) so that the Boltzmann factor can be linearized or expanded by retaining only the linear term without making too much error. In later analyses, δV will be represented by the lower-case v to conform with IEEE notation convention. The mathematical development of this expansion is made by decomposing the electric potential into a d.c. and a time-varying component,

$$V = V_{DC} + \delta V. \tag{410.1}$$

Then, the Boltzmann factor can be expanded by the Taylor series to give

$$\exp(qV/kT) = \exp[q(V_{DC}+\delta V)/kT] = \exp[qV_{DC}/kT]\cdot\exp[q\delta V/kT]$$

$$= \exp[qV_{DC}/kT]\cdot\{1 + (q\delta V/kT) + (q\delta V/kT)^2/2! + \dots\} \tag{410.2}$$

$$\simeq \exp[qV_{DC}/kT]\cdot\{1 + (q\delta V/kT)\}. \tag{410.2A}$$

The linearization condition to give (410.2A) is that the neglected higher order terms must be small compared with the retained linear term. This gives a quantitative criterion on the magnitude of the small-signal. For example, suppose that an accuracy of 1% is required. Then, the second order or quadratic term in the Taylor series of the exponential function given by (410.2) must be less than 1% of the first order or linear term, i.e.,

$$(q\delta V/kT)^2/2! \le (q\delta V/kT)/100 \tag{410.3}$$

or

$$\delta V \le (1/50)(kT/q) \tag{410.4}$$

$$\le 25.85\text{mV}/50 = 0.517\text{mV} \simeq 0.5\text{mV (at } T=300\text{K}). \tag{410.4A}$$

This result has a very important consequence. It shows that the small-signal requirement to linearize the system cannot be met by just requiring that the time-varying component of the potential or applied voltage is small compared with the applied d.c. voltage. It must also be small compared with the thermal voltage, kT/q. For example, if we were to apply the 1% criterion to the total or d.c. applied potential, $\delta V \le V_{DC}/100$, then the result would show that we could use a small-signal amplitude of $\delta V \le V_{DC}/100 = 5/100 = 50\text{mV}$ for $V_{DC}=5\text{V}$. This is equal to twice the value of the thermal voltage at room temperature, $kT/q \simeq 25\text{mV}$, and 100 times larger than true maximum for less than 1% error just computed, 0.5mV. For a $\delta V \simeq 50\text{mV}$, $\exp(\delta V/kT)$ would be highly nonlinear and give a highly nonlinear response (or highly nonlinear time-varying MOS capacitance current). The criterion developed in (410.4) is universal and not restricted to the MOSC or the specific numerical example just given. It is a unique electrical property of semiconductor devices which applies also to electrical transport in other solid state devices and material systems (biological and chemical) in which the charge carrier density is exponentially dependent on the potential energy via the Boltzmann factor.

The response of the MOSC to an applied small-signal perturbation comes from the movement or spatial redistribution of mobile charges (semiconductor electrons and holes). The capacitance does not vary with the small signal, It stays at the steady-state value as long as the applied d.c. voltage is kept constant during the application of the small signal to measure the capacitance. The small-signal equivalent circuit, Fig.410.1(d), is valid to analyze the response to both transient and sinusoidal steady-state signals as long as the signal amplitude is smaller than $(1/50)(kT/q) \simeq 0.5\text{mV}$.

The terminal small-signal capacitance of the MOSC is equal to the oxide capacitance in series with the semiconductor capacitance if all the oxide and interface charges do not change with time. This is simply given by

$$C_g = C_oC_S/(C_o + C_S). \qquad (410.5)$$

Subscript, g, denotes gate in conforming with the use of the MOS capacitor as the gate input electrode of the MOSFET. C_o is the oxide capacitance per unit area,

$$C_o = \varepsilon_o/x_o \qquad (410.6)$$

where ε_o is the static dielectric constant of the insulator. For SiO_2, ε_o=3.8x8.854x10^{-14} F/cm. x_o is the oxide or insulator thickness. To compute the total capacitance, C_g is to be multiplied by the gate area, A_g, which is the area of the metal or conductor gate electrode. The fringe electric field at the perimeter of the gate conductor electrode would give an effective area, A_g, slightly larger than the geometrical area, A_G.

The semiconductor capacitance, C_s, can be computed using the charge control analysis, $C_s = -dQ_S/dV_S$. V_S is the steady-state electric potential at the semiconductor surface or the oxide/silicon interface and known as the semiconductor surface potential. Q_S is the net space-charge density (C/cm^2) in the semiconductor. The series capacitance formulae was written down in (410.5) by intuitive circuit analysis. In fact, it can be derived using the charge control analysis as indicated by the following algebraic. From the space-charge distribution given in Fig.410.1(b),

$$C_g = dQ_G/dV_G = dQ_G/\{dV_O+dV_S\} \qquad (410.7)$$
$$= 1/\{(dV_O/dQ_G) + (dV_S/dQ_G)\} \qquad (410.7A)$$
$$= 1/\{(1/C_o) + [-dV_S/(dQ_S+dQ_{IT})]\} \qquad (410.7B)$$
$$= 1/\{(1/C_o) + [1/(C_s+C_{it})]\}. \qquad (410.7C)$$

From Kirchoff voltage law and Fig.410.1(c),

$$dV_G = dV_O + dV_S \qquad (410.8)$$

and

$$V_G = V_O + V_{BI} + V_S \qquad (410.9)$$

where V_{BI} is the built-in voltage (or potential drop) in the MOSC. C_o is defined by

$$C_o = dQ_G/dV_O. \qquad (410.10)$$

which is the oxide capacitance defined by (410.6). Q_S and Q_{IT} are respectively the charge density (C/cm^2) in the semiconductor and at the interface trap. These charges are related to the gate charge density, Q_G, shown in Fig.410.1(b), via the electrical neutrality condition

$$Q_G + Q_{OT} + Q_{IT} + Q_S = Q_G + Q_{OT} + Q_{IT} + Q_N + Q_P = 0 \qquad (410.11)$$

and

$$dQ_G + dQ_{IT} + dQ_S = dQ_G + dQ_{IT} + dQ_N + dQ_P = 0 \qquad (410.11A)$$

where dQ_{OT}=0 since since Q_{OT} is assumed a constant during the CV measurement.

In (410.7C), C_s is the differential semiconductor capacitance defined by

$$C_s = - dQ_S/dV_S \tag{410.12}$$

and C_{it} is the differential capacitance of the interface trap defined by

$$C_{it} = - dQ_{IT}/dV_S. \tag{410.13}$$

C_g of (410.7C) reduces to (410.5) if $C_{it}=0$ due to either Q_{IT}=constant or $Q_{IT}=0$.

The origin of the semiconductor capacitance, C_s, can be further dissected. It consists of two parts since there are two species of mobile charges, electrons and holes as indicated in Fig.410.1(b), and (410.11) and (410.11A). The two parts can be derived by computing the semiconductor areal charge density, Q_S, via integration of the semiconductor volume space-charge density, $\rho(x) = q(P - N - N_{AA})$ over the entire thickness of the semiconductor from $x=0$ to $x=+\infty$. We use the p-type semiconductor substrate indicated in Fig.410.1(a) and (b) as the model for this integration. It is doped uniformly with an acceptor impurity of N_{AA} ions per unit volume. We assume that all the acceptor impurity atoms are ionized so that $N_{AA}^- = N_{AA}$. Thus,

$$Q_S = \int_0^\infty q[P(x,V_S) - N(x,V_S) - N_{AA}(x)]dx \tag{410.14}$$

and

$$\begin{aligned}
C_s &= - dQ_S/dV_S \\
&= - (d/dV_S)\int_0^\infty q[P(x,V_S) - N(x,V_S) - N_{AA}(x)]dx && (410.15)\\
&= - (d/dV_S)\int_0^\infty \Big[q\{P(x,V_S)-P_B\} - q\{N(x,V_S)+N_B\}\Big]dx && (410.16)\\
&= \qquad - \{d(+Q_P)/dV_S\} - \{d(-Q_N)/dV_S\} && (410.17)\\
&= \qquad C_p + C_n. && (410.18)
\end{aligned}$$

In (410.15), we have used $N_{AA}=P_B-N_B$ which is a general relationship in the bulk region far away from the oxide/semiconductor and other contact interfaces. It was derived in chapter 2 from the charge neutrality condition at thermal and electrical equilibrium in the semiconductor bulk and was used to calculate the equilibrium concentration of the majority and minority carriers, P_B and N_B, with the subscript B, for bulk, omitted.

The semiconductor capacitance is now separated into a hole storage capacitance or hole charge-control capacitance C_p, and an electron storage capacitance or electron charge-control capacitance C_n, defined respectively by

$$C_P = -\, dQ_P/dV_S = -\,(d/dV_S)\int_0^\infty q[P(x,V_S) - P_B]dx \qquad (410.19)$$

and

$$C_n = -\, dQ_N/dV_S = -\,(d/dV_S)\int_0^\infty -q[N(x,V_S) - N_B]dx. \qquad (410.20)$$

The induced hole and electron charge density (C/cm^2) above are defined by

$$Q_P = +\int_0^\infty q[P(x,V_S) - P_B]dx \qquad (410.21)$$

and

$$Q_N = -\int_0^\infty q[N(x,V_S) - N_B]dx. \qquad (410.22)$$

Note that these induced carrier charge densities are those induced by (i) the d.c. voltage applied to the gate electrode, (ii) the imbedded oxide charge and charged interface traps, and (iii) the difference in electron affinity or work function of the three materials (metal, oxide, semiconductor). If $N_{AA}=f(x)$ or $N_{DD}=g(x)$, then there is a fourth contribution. If all of these charges are zero and the semiconductor is left at thermal and electrical equilibrium, shielded from light, and has no gate voltage applied, then the carrier concentrations are $P(x,V_S=0)=P_B$, $N(x,V_S=0)=N_B$, and there is no induced charges since $Q_P=Q_N=0$. When the gate voltage is positive, holes are repelled from the semiconductor surface layer under the gated oxide so that $Q_P < 0$, and electrons are attracted to the semiconductor surface layer under the gated oxide so that $Q_N < 0$. When the gate voltage is negative, holes are attracted to and electrons are repelled from the semiconductor surface layer so that $Q_P > 0$ and $Q_N > 0$. To shorten the notation, we have dropped the delta, Δ, in ΔQ_P and ΔQ_N whose retention would have eliminated the negative sign ambiguity in Q_P and positive sign ambiguity in Q_N.

The condition of $V_S=0$ which gives $Q_P=Q_N=0$ is also known as the **flat band** condition. It gives the value of the applied d.c. gate voltage that makes the energy band in the semiconductor flat or the electric field zero at the semiconductor surface. The semiconductor field may not be zero unless $N_{AA}=\text{constant}\neq f(x)$. The oxide field may not be zero if there is a distribution of charged oxide and interface traps, $\rho_{OT}(x)$ and Q_{IT}.

In the above analysis, we have made use of a very important fact which has eluded many transistor engineers and device physicists, that is, $N_{AA}(x)$ is fixed (independent of time) or $P_B(x)$ and $N_B(x)$ are both stationary. Hence, they cannot contribute to any capacitance. This cardinal point can be recognized from elementary physics: - differential capacitance is a measurement of the change of charge. Such a simple concept is frequently missed in textbooks and even research articles on semiconductor devices. For example, many authors have confused the depletion capacitance of a MOS capacitor and a p/n junction diode with the magnitude and spatial variation of the concentration of the fixed ionized impurities.

The depletion capacitance is a measure of the depleted mobile charge density and not the fixed ion concentration. The fixed ion concentration is measured only as a consequence provided the fixed ion concentration does not vary greatly with position, since then it is nearly equal to the majority carrier density that is measured by the depletion capacitance. This statement also gives the condition under which the depletion capacitance can measure the spatial variation of the impurity ion concentration. There is, however, a d.c. capacitance, $C_{DC} = Q_{DC}/V_{DC}$, where Q_{DC} does include the fixed charge density from the ionized impurities, but C_{DC} cannot be readily measured since Q_{DC} is not measured normally. This d.c. definition is in fact the Gauss theorem in disguise, which we have already used in (410.11) to relate the d.c. voltage to the d.c. charge. This d.c. relationship is used later to relate the d.c. surface potential, V_S, to the d.c. voltage, V_G, applied to the metal or conductor gate electrode.

The general integral of the two capacitances can be evaluated analytically. We make a change of variable from x to V(x) and use the boundary conditions of $V(0)=V_S$ and $V(\infty)=0$.

$$\begin{aligned} C_P &= -\,d(+Q_P)/dV_S = -\,(d/dV_S)\int_0^{\infty} q[P(x,V_S)-P_B]dx \\ &= -\,(d/dV_S)\int_{V_S}^{0} q[P(V,V_S)-P_B](dx/dV)dV \\ &= q[P(V_S)-P_B]/(dV/dx)\Big|_{V_S} = |q[P(V_S)-P_B]/(-E_S)| \qquad (410.23) \end{aligned}$$

$$C_n = -\,d(-Q_N)/dV_S = -q[N(V_S)-N_B]/(dV/dx) = |q[N(V_S)-N_B]/E_S|. \qquad (410.24)$$

where E_S is the d.c. electric field at the semiconductor side of the oxide/semiconductor interface, $x=0+$. It is given by $E_S=E(x=0+)=-dV(x)/dx$ at $x=0+$. The absolute sign is used since the two capacitances are always positive.

The capacitance due to charging and discharging the interface traps is more complicated. Its magnitude and voltage variation depend not only on the density but also on the energy level position of the interface trap. It is this variation that gives the distortion in the C_{LF} shown in Figs.403.1(c) and 403.2(c) which enables a measurement of the interface trap density by the Terman method [410.2]. This is ignored in the following analysis.

The results just obtained for C_p and C_n are given in terms of V_S and E_S. Thus, we need a second relationship to relate V_S and E_S to the applied d.c. gate voltage, V_G. This relationship can be easily derived from the Kirchoff's voltage law, i.e., the applied gate voltage is the sum of the potential drop across the oxide and semiconductor layers. Due to the imbedded oxide charge, charged interface traps, and work function difference between the gate conductor and the semiconductor, there will be a **built-in potential** drop across the oxide, V_{BIO}, and the semiconductor, V_{BIS}, when $V_G=0$. The Kirchoff's voltage law then gives

$$V_G = V_O - V_{BIO} + V_S - V_{BIS} \tag{410.25}$$

or $$V_G = V_O + V_S - V_{BI} \tag{410.26}$$

where $$V_{BI} = V_{BIO} + V_{BIS} \tag{410.27}$$

is the total built-in potential when $V_G=0$. It is the negative of the **metal/semiconductor work function difference**, $-\phi_{MS}$, which will be derived in section 412.

We must next find the dependence of V_O on V_S. This relationship can be derived from the charge neutrality condition given by (410.11)

$$Q_G + Q_{OT} + Q_{IT} + Q_S = 0. \tag{410.28}$$

Using the Gauss theorem, $Q_G = |\varepsilon_o E_o| = |\varepsilon_o dV_o/dx| = \varepsilon_o V_O/x_o = C_o V_O$ at the gate-conductor/oxide interface and $Q_S=-\varepsilon_s E_S$ at the oxide/silicon interface, (410.28) becomes

$$V_O = (\varepsilon_s E_S - Q_{OT} - Q_{IT})/C_o \tag{410.29}$$

which can then be used to replace V_O in (410.26) to give

$$V_G = [\phi_{MS} - (Q_{OT}+Q_{IT})/C_o] + V_S + \varepsilon_s E_S/C_o \tag{410.30}$$

$$= V_{FB} + V_S + \varepsilon_s E_S/C_o \tag{410.31}$$

where V_{FB} is the **flat-band voltage** defined by

$$V_{FB} = \phi_{MS} - (Q_{OT}+Q_{IT})/C_o. \tag{410.32}$$

It is the d.c. gate voltage required to flatten the energy band of the semiconductor, making $V_S=0$ and $E_S=0$. It measures directly the trapped oxide and interface charge densities.

The semiconductor capacitance from holes and electrons, C_p given by (410.23) and C_n given by (410.24), and the relationship between V_G and V_S, (410.31), are the most general results. They were derived based on elementary mathematics and electrostatics.

In addition to the above, one also needs a relationship between the semiconductor surface electric field, E_S, and semiconductor surface potential, V_S, in order to be able to compute the theoretical C-V curve, C_g vs V_G. The E_S-V_S relationship can be obtained from integrating the fifth Shockley equation, i.e., the Poisson equation. In the next section, 411, simple models are employed to give approximate analytical solutions of the Poisson equation. In section 412, energy band diagrams are used to obtain the exact solution.

411 Charge Control C-V Theory without Energy Band Diagram

In this section, we shall obtain the approximate analytical solutions of several C-V curves and C_g-V_G formulae using the **charge-control method** alone without using the energy band diagram. These include the depletion, high-frequency, low-frequency, and accumulation C-V curves; and the high-frequency flat-band (C_{fb}), intrinsic (C^i), strong-inversion threshold (C_{th}), and strong-inversion asymptotic (C_∞) capacitances. They are identified on the theoretical C-V curves given in Figs.411.1(a) and (b).

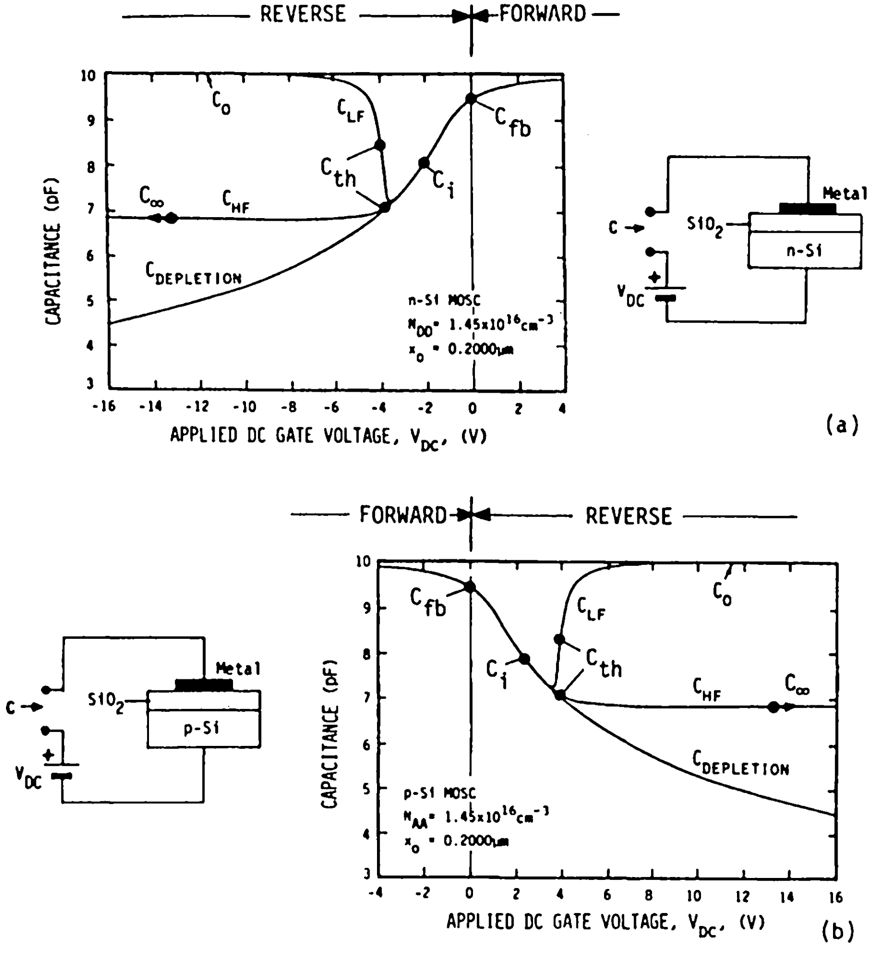


Fig.411.1 Theoretical capacitance-voltage curves of MOSC with analytical solutions obtained in the text labeled. (a) n-Si. (b) p-Si.

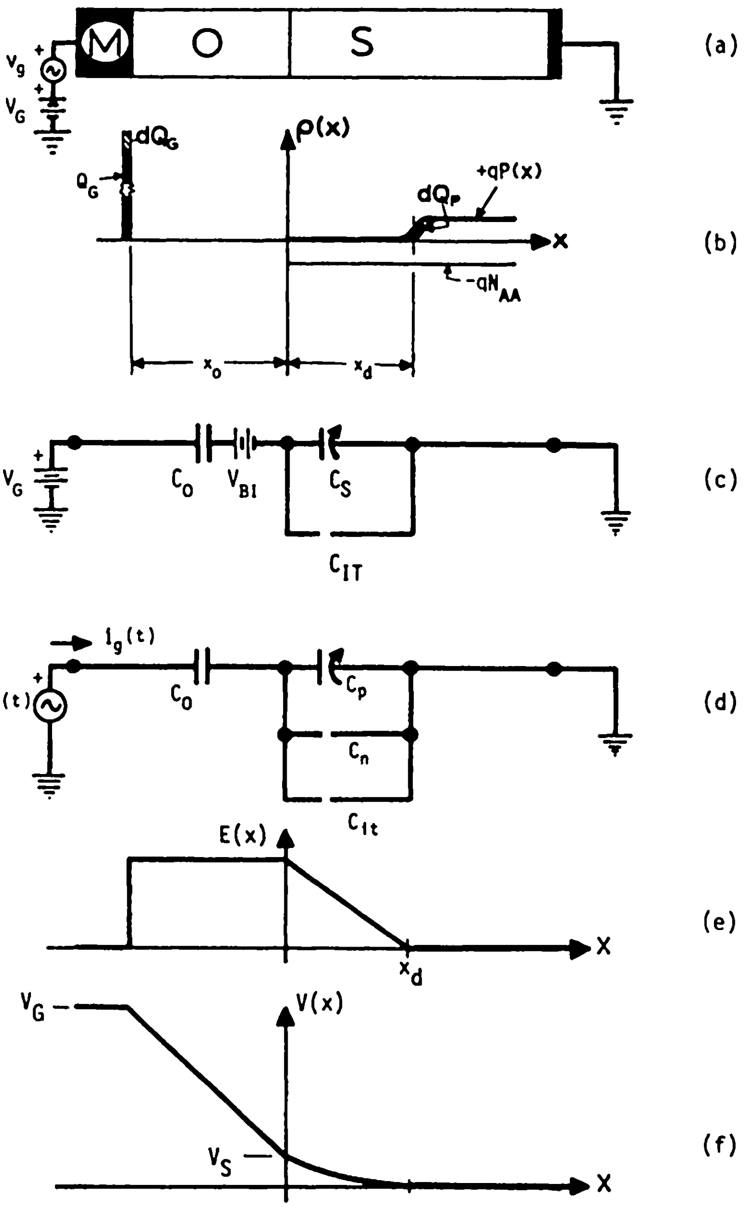

M
O
S
V_g
V_G
(a)
dQ_G
$\rho(x)$
Q_G
$+qP(x)$
dQ_P
X
$-qN_{AA}$
x_o
x_d
(b)
V_G
C_O
V_{BI}
C_S
C_{IT}
(c)
$I_g(t)$
$V_g(t)$
C_O
C_p
C_n
C_{It}
(d)
E(x)
X
x_d
(e)
V_G
V(x)
V_S
X
(f)

Fig.411.2 Depletion model of a p-Si MOSC. (a) Cross section. (b) Space-charge distribution. (c) d.c. and (d) small-signal equivalent circuits. (e) Electric field. (f) Electric potential. These figures are similar to those of Fig.410.1 except that here $Q_{IT}=0$ and $Q_N=0$.

(A) Depletion Capacitance

The analytical solution of the depletion case is the easiest to obtain because the electron and hole concentrations are assumed zero or the electrons and holes are depleted in the semiconductor surface space-charge layer under the gated oxide. Let the depletion layer be located between $x=0$ and $x=x_d$ under the gated oxide as indicated in Fig.410.1(b) which is redrawn in Fig.411.2(b) for the depletion model with the assumption $\rho_{OT}=0$ and $Q_{IT}=0$. $N(x)=0$ in the depletion model and also omitted in the figure. also shows that $N(x)=0$ and $P(x)=0$ in this depletion layer. Then the Poisson equation in the depletion layer is given by

$$\varepsilon_s dE/dx = -\varepsilon_s d^2V/dx^2 = \rho = q(P - N - N_{AA}) \qquad (411.1)$$

$$\approx q(-N_{AA}) \qquad 0 < x < x_d. \qquad (411.1A)$$

The boundary conditions are

$$E(x=x_d) = 0 \qquad (411.2A)$$

$$V(x=x_d) = 0 \qquad (411.2B)$$

and

$$V(x=0) = V_S. \qquad (411.2C)$$

Integrating the Poisson equation twice using these boundary conditions, then

$$V_S = qN_{AA}x_d^2/2\varepsilon_s \qquad (411.3)$$

or

$$x_d = \sqrt{2\varepsilon_s V_S/qN_{AA}} \qquad (411.3A)$$

$$E_S = qN_{AA}x_d/\varepsilon_s \qquad (411.4)$$

and

$$E_S = 2V_S/x_d = \sqrt{2qN_{AA}V_S/\varepsilon_s}. \qquad (411.5)$$

The electric field and potential obtained from the Poisson equation are shown in Fig.411.2(e) and (f). Note that the semiconductor electric field increases linearly towards the oxide/semiconductor interface and reaches the highest value at the interface, $x=0$. Substituting E_S into (410.30), we have

$$V_G = V_{FB} + V_S + \varepsilon_s E_S/C_o = V_{FB} + V_S + \sqrt{2\varepsilon_s qN_{AA}V_S}/C_o. \qquad (411.6)$$

which can be solved to give V_S as a function of V_G,

$$V_S = [\sqrt{(V_G - V_{FB}) + V_{AA}} - \sqrt{V_{AA}}] \qquad (411.6A)$$

where

$$V_{AA} = \varepsilon_s qN_{AA}/2C_o^2. \qquad (411.6B)$$

$C_n=0$ due to assuming depletion or $N=0$ in a p-Si MOSC. $C_{it}=0$ due to assuming $Q_{IT}=0$. Only $C_p\neq0$ even $P(x=0)=P_S=0$ inside the depletion layer because $P(x)$ rises to N_{AA} at the edge of the depletion layer, $x=x_d$, as indicated in Fig.411.2(b) and holes are moving in and out of the edge layer, dx_d, in an amount of dQ_P in response to time dependence of the applied signal voltage, dV_G, or current, dI_G.

Thus, C_p is given by the parallel-plate capacitance of the depletion layer ε_s/x_d. Using (411.3A) we obtain

$$C_s = C_n + C_p \simeq C_p \simeq C_d = \varepsilon_s/x_d = \sqrt{\varepsilon_s q N_{AA}/2V_S}. \tag{411.7}$$

$$= C_o[\sqrt{V_{AA}(V_G-V_{FB}) + V_{AA}^2}/(V_G-V_{FB})] \tag{411.7A}$$

The terminal capacitance from the series capacitance formulae, (410.5), is then

$$C_g = C_oC/(C_o+C_s) = C_o/\sqrt{1+[(V_G-V_{FB})/V_{AA}]} \tag{411.8}$$

which can be written in another rather useful form for transient analyses later

$$C_g^{-2} = C_o^{-2} + 2(V_G-V_{FB})/(\varepsilon_s q N_{AA}). \tag{411.8A}$$

At very large reverse gate bias, the voltage drop is mainly across the semiconductor surface space-charge layer, as indicated by (411.6), $V_G \simeq V_{FB}+V_S$. The semiconductor space-charger layer is so thick that $C_s << C_o$. Thus, the gate capacitance is mainly determined by C_s as if the oxide layer had thinned down to zero thickness. The capacitance is then given by

$$C_g \simeq C_s \simeq C_p \simeq C_d = \sqrt{\varepsilon_s q N_{AA}/2V_S} \simeq \sqrt{\varepsilon_s q N_{AA}/2(V_G-V_{FB})}. \tag{411.8B}$$

which can also be derived from (411.8A) directly. This is identical to the capacitance of a p+/n or metal/semiconductor (Schottky barrier) junction diode under reverse bias to be described in chapter 5. The depletion CV curves in Figs.402.1(a) and (b) and 411.1(a) and (b) show this $(V_G)^{-1/2}$ voltage dependence at large reverse bias. As a numerical illustration, consider a MOSC on p-Si with $N_{AA} = 10^{16}\text{cm}^{-3}$. Using $\varepsilon_s = 11.8 \times 8.854 \times 10^{-14}$F/cm, then $C_g = 1.3 \times 10^{-8}\text{F/cm}^2$ at $V_G - V_{FB} = 5$V. For a micron size MOSC with an area of $1 \times 1\mu\text{m}^2 = 10^{-8}\text{cm}^{-2}$, $C_gA_g = 1.3 \times 10^{-16}$F = 0.13fF.

(B) High-Frequency Capacitance

As indicated in Figs.411.1(a) and (b), the high-frequency capacitance of a MOSC drops to a gate-voltage independent low plateau asymptotic value, C_∞, when it is increasingly reverse biased. As explained in section 402, this constant plateau is caused by two factors: (i) the MOS insulator is leakless so minority carrier can accumulate at the oxide/semiconductor interface to invert the surface and (ii) the minority carrier density at the interface cannot change fast enough to respond to the signal voltage variation at 1MHz due to the lack of a fast enough and high enough density generation-recombination trap close to the interface). Consider the p-Si MOSC. In cause (i), the density of the minority carriers (electrons) accumulated at the oxide/semiconductor interface increases exponentially with surface potential, V_S, or applied gate voltage V_G, while the majority carrier (holes) depletion at the edge of the space-charge layer, x_d, increases only as $\sqrt{V_S}$ or $\sqrt{V_G}$. Consequently, when V_G is increased by an additional gate charge ΔQ_G, most of the ΔQ_G is used to attract electrons (minority carriers) to the surface inversion layer by increasing $|Q_N|$ (positive $|\Delta Q_N|$) while $|\Delta Q_P|$ is nearly zero or the edge of the space-charge layer, x_d, does not expand significantly. Thus, the dielectric capacitance of the space-charge layer $C_d=\varepsilon_s/x^d$ is nearly constant instead of decreasing continually with increasing V_G as predicted by (411.8B) for the depletion case. Because of the low minority carrier generation-recombination rate in cause (ii), the inversion electrons do not contribute to the high-frequency MOS capacitance. So the MOS high-frequency capacitance in strong inversion is just two constant capacitances in series, $C_g=C_oC_d/(C_o+C_d)$. These physics were first understood by Grove, Snow, Deal and Sah in 1962 [400.4]. They used the physics to derive a the correct and simple expression for the high-frequency capacitance which had not been derived by previous theorists because of complex mathematics. Their derivation is as follows using Figs.411.3(a)-(f).

Since the minority carrier density cannot change at the signal frequency, $dQ_N=0$ in Fig.411.3(b), thus, $C_n=0$ in Fig.411.3(d). Similarly, $C_{it}=0$ in Fig.411.3(d) either due to $Q_{IT}=0$ or low emission and capture rates at the interface traps so $dQ_{IT}/dV_S=0$. Then, C_p is the only remaining semiconductor capacitance as depicted in Fig.411.3(d). $C_p(V_G)$ is given by the depletion formula (411.7) in the surface potential range $0<V_S\lesssim 2V_F$. To calculate the capacitance at the threshold condition, $V_S=2V_F$, we use the definition: $N(x=0)=P_B=N_{AA}$ or lower $N(x=0)$ to N_{AA} and move x_{th} to $x=0$ in Fig.411.3(b). The gate voltage can be calculated using $V_S=2V_F$ in (411.6) and is given by

$$V_G = V_{FB} + V_S + \varepsilon_s E_S/C_o = V_{FB} + V_S + \sqrt{2\varepsilon_s q N_{AA} V_S}/C_{os} \qquad (411.6)$$

$$= V_{FB} + 2V_F + \sqrt{4\varepsilon_s q N_{AA} V_F}/C_o = V_{TH} = V_T \qquad (411.9)$$

where

$$V_F = (kT/q)\log_e(N_{AA}/n_i). \qquad (411.10)$$

This gate voltage is given the symbol V_{TH} or V_T which has been universally adopted by device as well as circuit engineers since 1964. It has a special significance in the operation of the MOS transistor, to be discussed in chapter 6: it is the minimum gate voltage required to strongly turn on the surface conduction channel or the transistor output current. Thus, it was coined the **threshold voltage** by this author in 1964 in the first paper on MOS transistor model [400.6]. The threshold point is labeled on both the HFCV and LFCV curves in Fig.411.1. At the threshold, the high-frequency semiconductor capacitance C_s from (411.7) is

$$C_S = \sqrt{\varepsilon_s q N_{AA}/2V_S} = \sqrt{\varepsilon_s q N_{AA}/4V_F}. \qquad (411.11)$$

At still higher d.c. gate voltage $V_G>V_{TH}$ and larger d.c. surface potential $V_S>2V_F$, most of the added ΔQ_G to the gate will attract additional electrons to the surface inversion layer, i.e., $C_o(V_G-V_{TH})=\Delta Q_G=\Delta Q_N+\Delta Q_P\simeq\Delta Q_N$. There is only a small additional depletion of holes, i.e. $|\Delta Q_P|<<|\Delta Q_N|$, which contributes to the slight further reduction of capacitance from $C_{th}(HF)$ to C_∞ shown in Figs.411.1(a) and (b). Accurate analytical theory worked out by this author in 1964 showed that this amounts to about 3kT/q additional V_S for $V_F = 10$ ($N_{AA} = 2.2\text{x}10^{14}\text{cm}^{-3}$ with 2.573kT/q) to $V_F = 16$ ($N_{AA} = 8.9\text{x}10^{16}\text{cm}^{-3}$ with 3.204kT/q), i.e. the drop to C_∞ is complete when $|\Delta Q_N|$ is $\exp(q\Delta V_S/kT) = e^3 = 20$ times larger than $|\Delta Q_P|$ for an additional ΔQ_G. So, $C_{s\infty}$ can be estimated from (411.11) with $2V_F$ replaced by $2V_F + 3(kT/q)$ and C_∞ can be calculated from $C_0C_{s\infty}/(C_0 + C_{s\infty})$. Note that this 3kT/q rise in V_S is reflected in potential and electric field spikes at x = 0 in Figs.411.3(e) and (f).

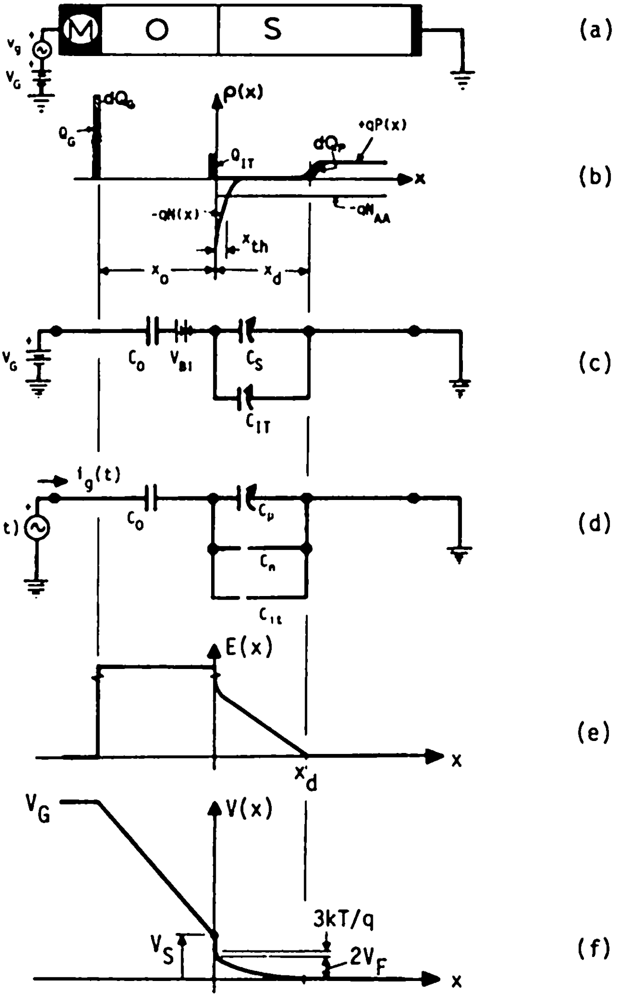

M
O
S
(a)
v_g
V_G
dQ_G
Q_G
ρ(x)
+qP(x)
dQ_P
Q_{IT}
x
-qN(x)
-qN$_{AA}$
x_{th}
x_o
x_d
(b)
V_G
C_o
V_{BI}
C_S
C_{IT}
(c)
$i_g(t)$
$v_g(t)$
C_o
C_p
C_n
C_{it}
(d)
E(x)
x
x'_d
(e)
V_G
V(x)
3kT/q
V_S
$2V_F$
x
(f)

Fig.411.3 High-frequency model of a p-Si MOSC. (a) Cross section. (b) Space-charge distribution. (c) d.c. and (d) small-signal equivalent circuits. (e) Electric field. (f) Electric potential.

(C) Low-Frequency Capacitance

The low-frequency capacitance at large reverse bias or strong inversion rises towards C_o as indicated by the CV curves in Figs.402.1(a) and (b), and 411.1(a) and (b). The reason of this rise was briefly discussed in section 402. To summarize, the large and increasing C_s with increasing inversion is due to two factors: (1) the sufficiently high rate of generation and recombination of the minority carriers to respond to the voltage variation at the very low signal frequency, and (2) the increasingly higher minority carrier density, Q_N in the semiconductor surface inversion layer of p-MOSC, as indicated by the -qN(x) curve in Fig.411.3(b). Thus, the assumption of $C_n=0$ in the preceding high-frequency capacitance analysis must be removed because dQ_N via generation-recombination can now follow the amplitude of the low-frequency signal. To compute C_n, a more accurate or nearly exact expression for the d.c. electric field E_S is needed in place of the depletion approximation, (411.5), which was used to derive the high-frequency capacitance.

To obtain E_S as a function of V_S without making the depletion approximation, we retain the carrier concentration in the space-charge density, $\rho(x)=q[P(x)-N(x)-N_{AA}]=q[P(x)-N(x)-P_B+N_B]$. For the inversion range in p-Si MOSC, $V_G\geq V_{TH}$, $V_S\geq 2V_F$, and $N(0)>P_B$. Also $P_B<<P(x)$ in the space-charge layer. Thus, we retain N(x) and drop P(x) and P_B. Although N_B is very small, it is retained so that the integral $N(x)-N_B$ is not divergent due to the integration limit at $x=\infty$ and to conform with the induced charge definition of Q_N given by (410.22). Using the Boltzmann relationship, $N(x)=N_B\exp(qV/kT)$, and the strong-inversion approximation for the minority carriers (electrons) just described, the Poisson equation becomes

$$\varepsilon_s dE/dx = -\ \varepsilon_s d^2V/dx^2 = \rho = q[P-N-P_B-+N_B] \simeq q[-N+N_B]$$
$$= qN_B[\exp(qV/kT)-1]. \qquad (411.12)$$

This can be integrated by quadrature, i.e.

$$d^2V/dx^2 = (dV/dx)(d/dV)(dV/dx)$$
$$= (1/2)(d/dV)(dV/dx)^2 = (1/2)dE^2/dV \qquad (411.13)$$

where $E=-dV/dx$ is the electric field. The integration constants can be evaluated using $V(x=0)=V_S>>0$, $V(x=\infty)=0$, $E(x=0)=E_S$ and $E(x=\infty)=0$. The result is

$$E_S = \sqrt{(2qN_B/\varepsilon_s)\{(kT/q)[\exp(qV_S/kT)-1]+V_S\}} \qquad (411.14)$$

$$\simeq \sqrt{(2kTN_B/\varepsilon_s)\exp(qV_S/kT)} \qquad (411.14A)$$

$$= \sqrt{(2kTN_{AA}/\varepsilon_s)\exp[q(V_S-2V_P)/kT]}. \qquad (411.14B)$$

The strong inversion approximation in (411.14A) is obtained when the gate voltage is sufficiently high that $\exp(qV_S/kT)>100(qV_S/kT)$ in the radical of (411.14), giving

$V_S > 6.5(kT/q)$. This is valid since the strong inversion condition already requires that $V_S > 2V_F$ which gives $V_S > 2V_F \simeq 20(kT/q)$ in extrinsic Si with $N_{AA} > 10^{14}cm^{-3} > 10^4 n_i$ which has a Fermi potential of $V_F = (kT/q)\log_e(N_{AA}/n_i) \geq 10kT/q$. Equation (411.14B) is obtained using $P_B = n_i\exp(qV_F/kT) \simeq N_{AA}$ and $N_B = n_i\exp(-qV_F/kT)$. This surface electric field can then be substituted into the gate voltage expression (410.31) to give

$$V_G = V_{FB} + V_S + \varepsilon_s E_S/C_o \tag{410.31}$$

$$\simeq V_{FB} + V_S + \varepsilon_s\sqrt{(2kTN_{AA}/\varepsilon_s)\exp[q(V_S-2V_F)/kT]}/C_o \tag{411.15}$$

$$\simeq V_{FB} + \sqrt{2\varepsilon_s kTN_{AA}\exp[q(V_S-2V_F)/kT]}/C_o. \tag{411.15A}$$

Approximation (411.15A) is valid when the second term in (411.15), V_S, is 10 times smaller than the third term. This gives the approximation criterion

$$V_S > 2V_F + 2(kT/q)\log_e[10(qV_S/kT)(\varepsilon_o/\varepsilon_s)(X_{DB}/\sqrt{2}x_o)] \tag{411.16}$$

$$> 2V_F + 4(kT/q) \tag{411.16A}$$

X_{DB} in (411.16) is the **extrinsic Debye length** defined by

$$X_{DB} = \sqrt{\varepsilon_s kT/q^2 N_{AA}} \tag{411.17}$$

It is the characteristic length of the exponential decay of a perturbation of the carrier density: $\delta p(x) = \delta p(0)\exp(-x/X_{DB})$. A mathematical demonstration will be given when the flat-band semiconductor capacitance is derived in subsection (E).

The $4(kT/q)$ in the inequality (411.16A) was obtained using a specific numerical example although it is valid over a range of parameters. To illustrate, consider a p-Si MOSC with $N_{AA} = 10^{16}cm^{-3}$, $\varepsilon_s = 11.8 \times 8.854 \times 10^{-14}F/cm$, $\varepsilon_o/\varepsilon_s = 4/12$ for SiO_2/Si, and $n_i = 10^{10}$ cm^{-3} at T=300K or kT/q=0.02585V. This gives a local Debye length of $290\sqrt{2} \simeq 410A$. Let the oxide thickness be 290A which is in the range used in early generations of Si VLSI circuits. The Fermi potential is $V_F = (kT/q)\log_e(N_{AA}/n_i) = 0.02585\log_e(10^{16}/10^{10}) = 13.81(kT/q) = 357mV$ or the Fermi level is 357meV below the intrinsic Fermi level. The minimum surface potential is then $V_S > 2V_F + 4(kT/q) = 31.63(kT/q) = 817.66mV$. The minimum gate voltage is then

$$V_G - V_{FB} > 10V_S = 20V_F + 40(kT/q) \tag{411.18}$$

$$\simeq 276(kT/q) + 40(kT/q) = 316(kT/q) = 8.17V. \tag{411.18A}$$

Under this strong inversion condition, the inversion layer capacitance, C_n, has a very simple dependence on the gate voltage. It can be derived by expressing the surface concentration of electron, N_S, and the electric field at the semiconductor surface, E_S, as a function of the gate voltage. These are

$$N_S = N_B\exp[qV_S/kT] = N_{AA}\exp[q(V_S-2V_P)/kT] \tag{411.19}$$

$$\approx C_o{}^2(V_G-V_{FB})^2/(2\varepsilon_s kT) \tag{411.19A}$$

$$E_S \approx \sqrt{(2kTN_{AA}/\varepsilon_s)\exp[q(V_S-2V_P)/kT]} \tag{411.14B}$$

$$\approx C_o(V_G-V_{FB})/\varepsilon_s. \tag{411.20}$$

Using these in C_n given by (410.24), then

$$C_n = |q(N_S-N_B)/E_S| \approx |qN_S/E_S| = [q(V_G-V_{FB})/2kT]C_o > 150C_o \tag{411.21}$$

Thus, this approximate solution is valid only when the inversion capacitance, C_n, is much larger than C_o or C_g is nearly C_o. To cover the less strongly inverted range, V_S in (411.15) cannot be dropped as we have in (411.15A).

(D) Accumulation Capacitance

In the accumulation range (V_G is negative in p-Si MOSC), the hole or majority carrier capacitance C_p dominates the semiconductor capacitance C_s. The results just obtained for the strong inversion range are all applicable if N_B is replaced by P_B and a change of the sign of V_S is made. This simple derivation without doing any algebra shows the power of a good choice of notation and the use of basic physics.

(E) Flat-Band Capacitance

This is an interesting case that contains important solid state physics. Since the flat-band is defined as $V_S=0$ and $E_S=0$ in the semiconductor, or the semiconductor energy band is horizontal, we have

$$P_S-P_B = N_S-N_B = Q_P = Q_N = E_S = 0. \tag{411.22}$$

These zeros make C_p and C_n defined by (410.23) and (410.24) indeterminant, 0/0. One approach to obtain the asymptotic expressions is to grind through the mathematics by evaluating the numerator and denominator in the limit of $V_S \to 0$. Instead of this brute force route, it is more satisfactory to make the necessary approximations at the starting point of the analysis since it would give us new physical insights on semiconductor and device properties.

At flat band, $V_S=0$, the small-signal voltage v_g will produce a small-signal variation of the electron and hole charge densities at the oxide/semiconductor interface. This variation will decay with distance exponentially with a characteristic length known as the Debye length at the semiconductor surface. It is most easily derived by expanding the electron and hole concentrations in Taylor series since δV_S is much smaller than kT/q. We denote the small-signal variation of the electric potential, δV, by v_i (i from intrinsic Fermi potential). Thus,

$$P(x) = P_B\exp(-qV/kT) = P_B\exp(-q\delta V/kT) = P_B\exp(-qv_i/kT)$$

$$\simeq P_B[1 - qv_i/kT] \qquad (411.23)$$

and

$$N(x) = N_B\exp(+qV/kT) = N_B\exp(+q\delta V/kT) = N_B\exp(+qv_i/kT)$$

$$\simeq N_B[1 + qv_i/kT]. \qquad (411.24)$$

These can be substituted into the Poisson equation to give

$$\begin{aligned} \varepsilon_s dE/dx &= -\ \varepsilon_s d^2V/dx^2 = -\ \varepsilon_s d^2(\delta V)/dx^2 \\ &= -\ \varepsilon_s d^2v_i/dx^2 \\ = \ \rho &= q(P - N - N_{AA}) \\ &= q[P_B - P_B(qv_i/kT) - N_B - N_B(qv_i/kT) - N_{AA}] \\ &= q[-q(P_B+N_B)/kT]v_i \qquad (411.25) \end{aligned}$$

where we have used the d.c. charge neutrality condition, P_B-N_B-$N_{AA}=0$. Thus, the small-signal Poisson equation at flat-band is linearized and given by

$$d^2v_i/dx^2 = [q^2(P_B+N_B)/\varepsilon_s kT]v_i = v_i/L_{DB}{}^2 \qquad (411.26)$$

where L_{DB} is the local Debye length defined by

$$L_{DB} = \sqrt{\varepsilon_s kT/q^2(P_B+N_B)} \qquad (411.26A)$$

which reduces to the extrinsic local Debye length defined by (411.17). Equation (411.25) can be integrated by quadrature using $d^2v_i/dx^2 = (1/2)(d/dv_i)(dv_i/dx)^2$. The integration constants are evaluated using the boundary conditions $v_i(x=0)=v_s$, $v_i(x=\infty)=0$, $dv_i(x=0)/dx=-E_s$, and $dv_i(x=\infty)/dx=0$. The result is an exponential decay of the signal potential, v_i, away from the interface,

$$v_i(x) = v_i(0)\exp(-x/L_{DB}) \qquad (411.26B)$$

The small-signal electric field is then $E_s = -dv_i(x=0)/dx = -v_s/L_{DB}$. Using P_S-P_B $= P_B[\exp(-qv_s/kT)-1] \simeq -P_B(qv_s/kT)$, N_S-$N_B = N_B[\exp(qv_s/kT)-1] \simeq N_B(qv_s/kT)$, and $E_s=-v_s/L_{DB}$ in (410.23) and (410.24), the capacitances at flat-band are

$$C_{pfb} = q^2P_BL_{DB}/kT \qquad (411.27)$$

and

$$C_{nfb} = q^2N_BL_{DB}/kT \qquad (411.28)$$

$$C_{sfb} = C_{pFB} + C_{nFB} = \varepsilon_s/L_{DB}. \qquad (411.29)$$

The last expression is particularly revealing. It states that the semiconductor capacitance at flat band is the dielectric capacitance of a semiconductor layer of a dielectric constant ε_s and a thickness equal to the local Debye length of the carriers in the semiconductor, L_{DB}. It implies that the small-signal charge propagates to an effective depth equal to the Debye length. At 300K, $L_{DB}=410$A at $N_{AA}=10^{16}$cm^{-3} in Si as calculated after (411.17) while the intrinsic Debye length in Si is 29.05μm using $P_B=N_B=n_i$ in (411.26A) which is obviously not a very useful concept. (See problem P411.11.)

(F) Summary

The results obtained in the preceding subsections, (A) to (E), are labeled on the three theoretical C-V curves for the n-Si and p-Si MOSC's in Fig.411.1(a) and (b). A summary of the analytical solutions and a description of applications to measure material parameters are given as follows.

The depletion capacitance at large reverse bias voltages, from (411.8B), is

$$C_{gd} \simeq C_s \simeq C_p \simeq \sqrt{\varepsilon_s q N_{AA}/2V_S} \simeq \sqrt{\varepsilon_s q N_{AA}/2|V_G-V_{FB}|} \qquad (411.30)$$

which can be used to determine N_{AA} experimentally. An improved formula is given by (411.8A) or (411.8).

The asymptotic high-frequency capacitance at large reverse bias or strong inversion is given by (411.11) with V_S set to $2V_F+3(kT/q)$. Thus, the high-frequency MOSC capacitance at strong inversion is

$$C_{g\infty} = C_oC_{s\infty}/(C_o+C_{s\infty}) = C_o/[1+C_o/\sqrt{\varepsilon_s q N_{AA}/2(2V_F+3kT/q)}]. \qquad (411.31)$$

This can also be used to determine the N_{AA} experimentally since V_F is given by $V_F=(kT/q)\log_e(N_{AA}/n_i)$. If N_{AA} is accurately known, then it can even be used to determine V_F or n_i.

The high-frequency capacitance at the onset of strong inversion is known as the threshold capacitance and the gate voltage is known as the threshold voltage. These correspond to $V_S=2V_F$ and are given by (411.11) and (411.9) respectively

$$V_{TH} = V_{FB} + 2V_F + \sqrt{4\varepsilon_s q N_{AA} V_F}/C_o \qquad (411.32)$$

and

$$C_{th} = C_oC_s/(C_o+C_s) = C_o/[1+C_o/\sqrt{\varepsilon_s q N_{AA}/4V_F}] \qquad (411.33)$$

where

$$V_F = (kT/q)\log_e(N_{AA}/n_i). \qquad (411.33A)$$

At intrinsic surface, $P(x=0)=N(x=0)=n_i$ and $V_S=V_F$. So, the gate voltage and the MOSC capacitance can again be obtained by the high-frequency or depletion formulae (411.6) and (411.7) using $V_S=V_F$. These are

$$V_{GI} = V_{FB} + V_F + \sqrt{2\varepsilon_s q N_{AA} V_F}/C_o \qquad (411.34)$$

and

$$C_i = C_oC_{si}/(C_o+C_{si}) = C_o/[1+C_o/\sqrt{\varepsilon_s q N_{AA}/2V_F}]. \qquad (411.35)$$

The flat-band capacitance is independent of frequency since it is dominated by majority carriers. It is given by the semiconductor Debye capacitance (411.29)

$$C_{fb} = C_oC_{sfb}/(C_o+C_{sfb}) = C_o/\{1+C_o/[\varepsilon_s/\sqrt{\varepsilon_s kT/q^2(P_B+N_B)}]\}. \qquad (411.36)$$

412 Advanced Charge-Control C-V Theory

We have derived the CV curves in the preceding section by piece-wise approximations in order to bring out the physics at each step and in each bias range. These can all be combined into one analysis if we retain both the electron and hole concentrations while integrating the Poisson equation. We shall present the algebra without repeating the detailed discussion of the underlying physics which was given in the preceding section. These analytical results have been used to analyze the experimental CV curves for monitoring the fabrication processing steps in Si integrated circuit production and for fundamental studies of the reliability physics and failure mechanisms of Si MOS transistors,

(A) Relating Electric Field to Potential at Semiconductor Surface

The d.c. steady-state Poisson equation

$$\begin{aligned} \varepsilon_s(dE/dx) &= -\ \varepsilon_s d^2V/dx^2 \\ &= \rho(x) = q[P(x) - N(x) - N_{AA}] \\ &= q[P(x) - N(x) - P_B + N_B] \end{aligned} \qquad (412.1)$$

is integrated by quadrature using E = - (dV/dx) to give,

$$\begin{aligned} \varepsilon_s(dE/dx) &= \varepsilon_s(d/dx)(-dV/dx) = -\ \varepsilon_s(dV/dx)(d/dV)(dV/dx) \\ &= -\ (\varepsilon_s/2)(d/dV)(dV/dx)^2 \\ &= -\ (\varepsilon_s/2)(dE^2/dx). \end{aligned} \qquad (412.2)$$

Then using the Boltzmann relationship for the carrier concentrations

$$P(x) = P_B\exp(-qV/kT) \qquad (412.3A)$$

and

$$N(x) = N_B\exp(+qV/kT) \qquad (412.3B)$$

in the macroscopic space-charge density of the Poisson equation, then

$$\begin{aligned} \varepsilon_s dE/dx &= -\ (\varepsilon_s/2)(d/dV)E^2 \\ &= q[P_B\exp(-qV/kT) - P_B] - q[N_B\exp(+qV/kT) - N_B] \\ &= qP_B[\exp(-qV/kT) - 1]\ -\ qN_B[\exp(+q/kT) - 1]. \end{aligned} \qquad (412.4)$$

The terms in (412.4) are put into two groups to emphasize their physical origin. The first two terms comes from the holes and the second, from electrons.

The Poisson equation can now be integrated by quadrature as follows.

$$\int_{V_S}^{0} \varepsilon_s(d\mathcal{E}/dx)dV = \int_{V_S}^{0} \varepsilon_s(d/dx)(-dV/dx)dV = \int_{V_S}^{0} \varepsilon_s(dV/dx)d(-dV/dx)$$

$$= \left.(\varepsilon_s/2)(dV/dx)^2\right|_{V_S}^{0} = -\varepsilon_s\mathcal{E}_S{}^2/2$$

$$= +\int_{V_S}^{0} \{qP_B[\exp(-qV/kT) - 1] - qN_B[\exp(+qV/kT) - 1]\}dV$$

$$= +qP_B\{(kT/q)[\exp(-qV_S/kT) - 1] + V_S\}$$

$$+ qN_B\{(kT/q)[\exp(+qV_S/kT) - 1] - V_S\}. \qquad (412.5)$$

Thus, the electric field at the semiconductor surface or the oxide/semiconductor interface, $x=0$ and $V(x=0)=V_S$, is given by

$$\mathcal{E}_S = \sqrt{2/\varepsilon_s}\cdot$$

$$\sqrt{qP_B\{(kT/q)[\exp(-qV_S/kT)-1]+V_S\} + qN_B\{(kT/q)[\exp(+qV_S/kT)-1]-V_S\}} \qquad (412.6)$$

$$= -(\mathrm{sign}V_S)\sqrt{2kT/\varepsilon_s}\cdot\sqrt{\{(P_S-P_B) + (qV_S/kT)P_B\} + \{(N_S-N_B) - (qV_S/kT)N_B\}}. \qquad (412.7)$$

The sign of the electric field is the negative of the sign of the surface potential, V_S. This d.c. electric field expression can be put into a more illustrative form as a ratio of a potential divided by a distance:

$$\mathcal{E}_S = -(\mathrm{sign}V_S)\sqrt{2kT/\varepsilon_s}\cdot\sqrt{\{(P_S-P_B)+(qV_S/kT)P_B\} + \{(N_S-N_B)-(qV_S/kT)N_B\}}$$

$$= -(\mathrm{sign}V_S)[(kT/q)/L_{DB}] \qquad (412.8)$$

where L_{DB} is the d.c. **Local Debye Length**. In this expression, it is the value at the semiconductor surface or the oxide/semiconductor interface and is defined by

$$L_{DS} = \left[\frac{(\varepsilon_s kT/2q^2)}{\{(P_S-P_B) + (qV_S/kT)P_B\} + \{(N_S-N_B) - (qV_S/kT)N_B\}}\right]^{1/2}. \qquad (412.9)$$

The electric field expression, just derived from elementary calculus, appears to be complicated at first glance but is not really if we identify the physical origin of the four terms under the radical. The two terms containing bulk concentration of holes, P_B, come from the influence of the applied d.c. gate voltage on the d.c. steady-state hole distribution near the surface through the variation of the surface potential V_S with gate voltage, V_G. When the surface potential is large and negative or the applied voltage is large and negative, holes are attracted to the oxide/semiconductor interface making the hole concentration near the oxide/semiconductor interface very high. This is obvious because the term $\exp(-qV_S/kT)$ would dominate when V_S is very negative.

Similarly, the last two terms containing N_B under the radical in (412.6) come from the electrons. The term, $N(x=0)=N_S=N_B\exp(+qV_S/kT)$, would dominate if the voltage applied to the gate is positive, making V_S large and positive. This is expected since the positive voltage would attract a high concentration of electrons to the oxide/semiconductor interface, while repelling holes into the interior far away from the interface.

(B) Relating Surface Potential to Gate Voltage

We need a second relationship that relates the surface potential, V_S, to the applied d.c. gate voltage, V_G, in order to obtain the dependence of the semiconductor surface electric field, E_S, as a function of applied d.c. gate voltage. This is obtained from two relationships, Kirchoff's Voltage Law and Gauss's Theorem. Two relationships are needed because both V_S and V_O varies with V_G. Kirchoff's law gives

$$V_G = V_O + V_S - V_{BI} \qquad (412.10)$$

where V_O and V_S are the potential drops across the oxide and semiconductor respectively and V_{BI} is the built-in potential drop across the entire MOSC when the applied voltage is zero, $V_G=0$, i.e.,

$$V_{BI} = V_O(V_G=0) + V_S(V_G=0). \qquad (412.11)$$

The built-in potential, V_{BI}, is a function of the material properties of the three layers: the metal or conductor gate, the oxide insulator layer, and the semiconductor substrate. To obtain this functional dependence, the E-x energy band diagram of the MOSC shown in Fig.412.1(c) must be used. A detailed step-by-step derivation of the E-x diagram will be given in the next section, 413. The symbols and features of this MOSC E-x energy band diagram will be described in the following paragraphs, starting from the left towards the right of Fig.412.1(c).

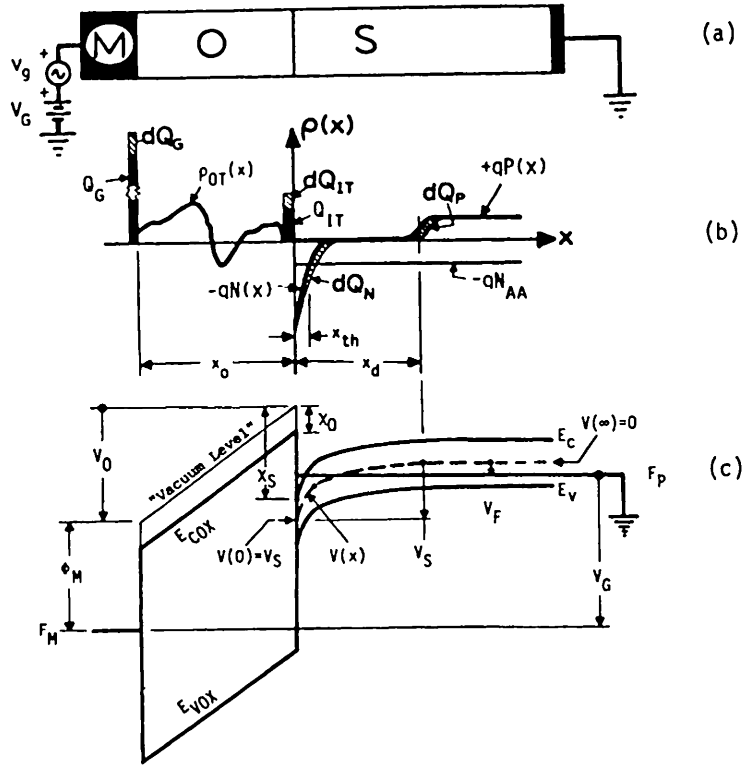


Fig.412.1 The E-x energy band diagram of a MOS capacitor on p-Si. (a) The cross-sectional view, the d.c. bias voltage and the a.c. test signal generator. (b) The space-charge distribution. The oxide charge, Q_{OT}, is shown in the illustration but not included in the constructing the oxide conduction and valence band edges, E_{Cox} and E_{Vox}, which have constant slope. (c) The E-x energy band diagram.

F_M is the Fermi level of the metal and F_P is that of the semiconductor. They are measured from an unspecified common reference energy level. ϕ_M is the work function of gate conductor or gate metal. V_O is the potential drop across the oxide. E_{Cox} is the oxide conduction band edge and E_{Vox} is the oxide valence band edge. The energy gap of silicon dioxide is about 8 eV, i.e., $E_{Cox}-E_{Vox}=8.0$eV. The oxide band edges are shown as straight sloped lines to indicate a constant oxide electric field. The straight lines from constant oxide electric field are used for simplicity of illustration and they do not take into account the oxide charge distribution $\rho_{OT}(x)$ pointed by the Q_{OT} arrow in Fig.412.1(b) which would distort the straight line significantly. The "vacuum level" above the oxide conduction band is the position of the vacuum level if the oxide, the conductor gate and the semiconductor were all far apart instead of in intimate contact as shown. It is meaningless in the infinite area oxide film sandwiched between the metal gate and the semiconductor since the space is occupied by the oxide and is not a vacuum-void. Nevertheless, it is retained to illustrate χ_o (the electron affinity of the oxide), χ_s (the electron affinity of the semiconductor), and ϕ_M (the work function of the metal). Note that the electron affinity is the energy measured from the bottom edge of the conduction band to the vacuum level at the surface of the insulator (oxide) or semiconductor. V(x) is the electric potential in the semiconductor which is defined by $V(x)=E_I(x)/(-q)$, i.e., the intrinsic Fermi level is used for the electric potential. This choice is arbitrary, since the reference potential is arbitrary. $V(x=0)=V_S$ is the electric potential at the semiconductor surface or at the oxide/semiconductor interface which is located at x=0 in our choice of coordinate system. It is known as the **surface potential**. $V(x=\infty)=0$, that is, the reference for the semiconductor electric potential is selected as a point far to the right from the oxide/semiconductor interface. A semi-infinite thick semiconductor is assumed. In practice, the thickness is finite, about 500 μm, and the mathematical boundary condition of $V(x=\infty)=0$ implies that the thickness of the semiconductor is many times the largest characteristic length associated with the signal charge propagation through the semiconductor, such as the diffusion length, the Debye length, and the depletion layer thickness. V_F is the Fermi potential in the bulk of the semiconductor measured from the intrinsic Fermi potential, $V_I=V(x=\infty)=0$. For the p-Si illustrated in Fig.412.1(c), $V_F = (kT/q)\log_e(N_{AA}/n_i)>0$ and for n-Si, $V_F = (kT/q)\log_e(n_i/N_{DD})<0$, from the analyses in chapter 2. F_P is the Fermi level in the bulk region of the p-type semiconductor measured from the same reference energy level as the Fermi level of the metal, F_M. 'Bulk' means the region at a large distance from the oxide/semiconductor and back contact interfaces.

From the energy diagram in Fig.412.1(c), we can immediately write down the following equation for the potential using Kirchoff's voltage law.

$$\phi_M + V_O = \chi_s - V_S + (E_C-F_P)/q + V_G \tag{412.12}$$

or

$$V_G = V_S + V_O + \phi_M - \chi_s - (E_C-F_P)/q \tag{412.12A}$$

$$= V_S + V_O + \phi_M - \chi_s - (E_C-E_I)/q - (E_I-F_P)/q \tag{412.12B}$$

$$= V_S + V_O + \phi_M - \chi_s - (E_C-E_I)/q - V_F \tag{412.12C}$$

$$\equiv V_S + V_O + \phi_M - \phi_S \tag{412.12D}$$

$$\equiv V_S + V_O + \phi_{MS} \tag{412.12E}$$

where $\phi_{MS}\equiv\phi_M-\phi_S$ is known as the **metal-to-semiconductor work function difference**. The result, (412.12E), is identical to the intuitive relationship (412.10). However, the intuitive built-in potential, V_{BI}, is now proven to be the negative of the metal-to-semiconductor work function difference, and explicitly related to the material parameters by

$$\phi_{MS} = \phi_M - \phi_S = \phi_M - \{\chi_s + [(E_C-E_I)/q] + V_F\}. \tag{412.12F}$$

This shows that the metal/semiconductor work function difference is determined by the fundamental material parameters of the metal and the semiconductor. And it is not affected by the intervening oxide layer and is not affected by the oxide and interface traps.

The result given by (412.12E) or (412.10) contains two unknowns, V_O and V_S as stated after (412.10). Thus, another relationship between V_O and V_S is necessary to give the functional dependence of V_S on V_G. This is obtained by the d.c. or non-differential (integrated) Gauss theorem. The differential Gauss theorem across the MOSC was used earlier to derive an expression for the semiconductor capacitance, C_s, as a function of the surface potential of the semiconductor at the oxide/silicon interface, V_S. The differential Gauss theorem was used since the small-signal semiconductor capacitance is a measure of the change of stored charge due to a change of potential, $C_s=dQ_S/dV_S$. The non-differentiated Gauss theorem applied across all MOSC layers, including the two contact metal layers or bodies, is now used here. This gives

$$\int_{-x_o}^{\infty} \rho(x)dx = 0 = Q_M + Q_{OT} + Q_{IT} + Q_S \quad . \qquad (412.13)$$

$$= C_oV_O + Q_{OT} + Q_{IT} - \varepsilon_s E_S . \qquad (412.13A)$$

In (412.13A), use is made of $Q_M=C_oV_O$ and $Q_S+D_S(x=0)=Q_S+\varepsilon_s E_S(x=0)=0$. Both of which can be obtained also from the Gauss theorem, the first from Gauss theorem applied to the metal-gate/oxide interface and the second, the semiconductor layer. Here as previously, Q_{OT} is not the integrated volume density of oxide charge, $\int \rho_{OT}(x)dx$, through the oxide layer but an effective oxide charge density located at the oxide/semiconductor interface which would give the same total potential drop across the semiconductor as the distributed oxide charge, $\rho_{OT}(x)$, in the oxide layer. It is given by

$$Q_{OT} = \int_0^{x_o} (x/x_o)\rho_{OT}(x)dx \qquad (412.13B)$$

where x_o is the oxide thickness. This formulae shows that x_oQ_{OT} is the moment of the oxide charge density, $<x\rho_{OT}>$, while Q_{OT} is the normalized moment, $<x\rho_{OT}>/x_o$. Equation (412.13B) can readily be derived by the parallel plate capacitance formulae applied to the volume distribution of charges in the oxide and is left as an exercise. A more tedious derivation is to integrate the Poisson equation through the entire MOS structure as show in (412.13).

Rearranging the solution above, the oxide potential drop is then

$$V_O = (\varepsilon_s E_S - Q_{OT} - Q_{IT})/C_o . \qquad (412.14)$$

Substituting this expression for V_O in (412.12E) and using the expression of E_S as a function of V_S derived in (412.8), we then have the final answer: a transcendental equation for V_S as a function of V_G given by

$$V_G = \phi_{MS} - (Q_{IT}+Q_{OT})/C_o + V_S + \varepsilon_s E_S/C_o \tag{412.15}$$

$$= V_{FB} + V_S + \varepsilon_s E_S/C_o \tag{412.15A}$$

where E_S is given by (412.8) and the flat-band voltage is defined by

$$V_{FB} = \phi_{MS} - (Q_{IT}+Q_{OT})/C_o. \tag{412.15B}$$

(C) The Exact Low-Frequency MOS Capacitance

The complete and exact solution for the low frequency capacitance as a function of the applied gate voltage is obtained using the semiconductor capacitances from (410.23) and (410.24)

$$C_s = C_p + C_n = q(|P_S-P_B|+|N_S-N_B|)/|E_S| \tag{412.16}$$

and the electric field expression from (412.6)-(412.9). This complete solution enables us to compute the MOS capacitance as a function the applied d.c. voltage, V_G, using the parallel-plate capacitance formulae, (410.5).

(D) Depletion and High-Frequency Capacitances

To obtain the depletion and high-frequency capacitances, the semiconductor capacitance formulae just obtained must be modified while the d.c. electric field and d.c. surface potential formula are still valid. For the high-frequency case, the semiconductor capacitance for a p-Si MOSC is defined as

$$C_s = -\left.dQ_S/dV_S\right|_{\substack{Q_N=\\ \text{constant}}} = -\left.dQ_P/dV_S\right|_{\substack{Q_N=\\ \text{constant}}} = C_p \tag{412.17}$$

which means that

$$C_n = -dQ_N/dV_S = 0 \tag{412.18}$$

must be substituted into Q_P first before the differentiation of Q_P is carried out. Various approximation schemes have been developed to evaluate the integral given by (412.17) which are described in an advanced course and in the advanced topic sections of chapter 6 on MOST.

For the depletion case, we may discard Q_N and retain only the majority carriers, P, P_B, and Q_P. The result is identical to that obtained for the accumulation capacitance.

413 Energy Band Diagram of MOSC

The E-x energy band diagram of the MOSC shown in Fig.412.1(c) was used in the preceding section, 412, to relate the built-in potential to the properties of the materials and to obtain the exact analytical solution of the electric field as a function of the surface potential or total semiconductor energy band bending, V_S. The E-x diagram can be derived rigorously using a step-by-step procedure which is illustrated in Fig.413.1(a) to (e). To proceed with the derivation, we assume that the three materials are at thermal equilibrium (at the same temperature) with a heat sink (the ambient air molecules). We also assume that initially they are not in electric contact and there is no electron current flowing between them since there is no pathway (or interconnection conductors) for the electrons to flow between them. Nevertheless, they could be in quasi-electrical equilibrium. The quasi condition can come about if they are charged to different potentials initially. Then, their potentials are quasi-static (does not change in a short observation time but does change during a long observation period) because the collisions of the ambient molecules at the solid surfaces that maintain thermal equilibrium will transfer some charges, however minute, among the three materials. This slow charge transfer will eventually equalize the potential on the three bodies to equilibrate them electrically also while they have already been in thermal equilibrium due to the frequent random molecular collision on the surface. We speed up the approach to electrical equilibrium by contacting the three bodies and connecting the Al and the p-Si electrically with a conductor of low electrical resistance as illustrated in the Figs.413.1(a) to (e) and explained in the following paragraphs.

Figures 413.1(a) shows the E-x energy band diagrams of the Al-metal, SiO_2, and p-Si when they are isolated and not in electrical contact with each other. In metal, the valence electrons fill up the conduction band levels partially to an energy F_M below the vacuum level. The metal Fermi level is in the conduction band. Its position distinguishes a metal (very high electrical conductivity) from a semiconductor (medium to low electrical conductivity) whose Fermi level is in the energy gap. The electron Fermi level in a metal measured from the bottom edge of the metal conduction band can be computed using the nearly-free electron energy band model, $E=\hbar^2k^2/2m$, the Fermi occupation factor, $f(E)=1/\{1+\exp[(E-E_F)/kT]\}=1$ for $E\leq E_F$ and $f(E)=0$ for $E>E_F$, and the valence electron concentration, N_v, obtained from the metal atomic density and the valence electron number. This is given by

$$E_F-E_C = (\hbar^2/2m)(3\pi^2N)^{2/3} \tag{413.1}$$

$$= (2.202\times10^{-22}N_v)^{2/3} \text{ (eV)} \tag{413.1A}$$

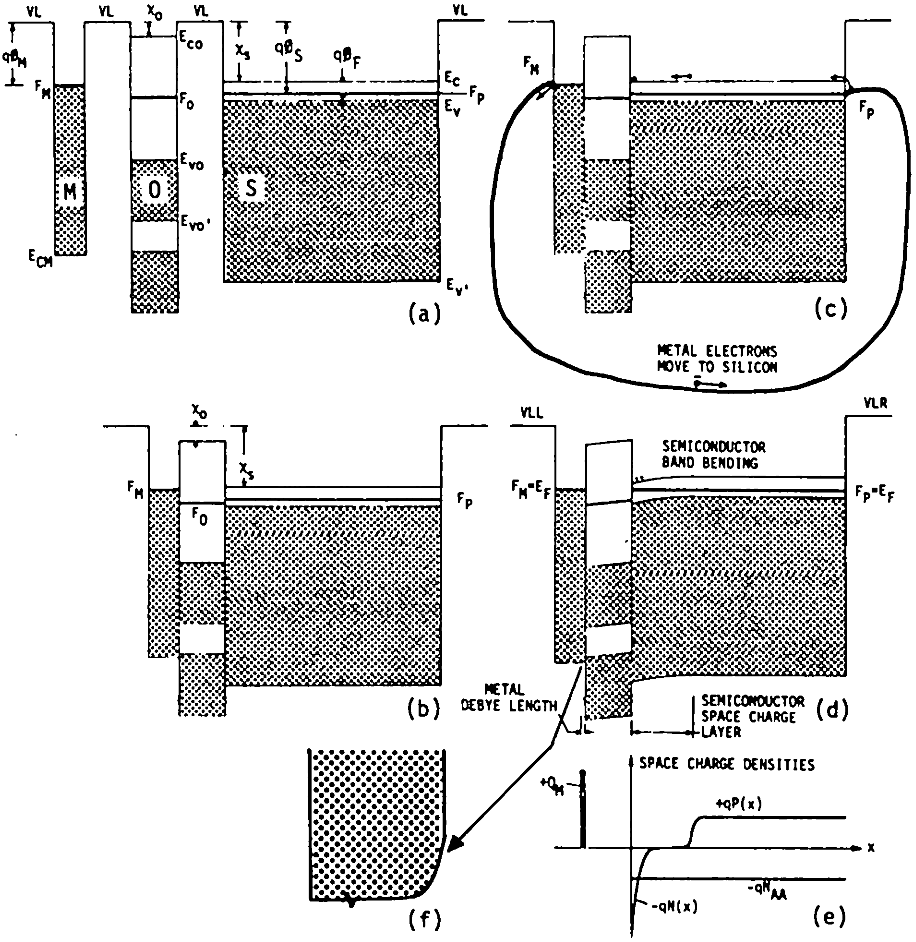

VL
VL
X_O
VL
VL
$q\phi_M$
F_M
E_{CO}
X_S
$q\phi_S$
$q\phi_F$
E_C
F_P
E_V
F_O
E_{VO}
$E_{VO'}$
M
O
S
E_{CM}
$E_{V'}$
(a)
F_M
F_P
(c)
METAL ELECTRONS MOVE TO SILICON
X_O
X_S
F_M
F_O
F_P
(b)
VLL
VLR
SEMICONDUCTOR BAND BENDING
$F_M=E_F$
$F_P=E_F$
METAL DEBYE LENGTH
SEMICONDUCTOR SPACE CHARGE LAYER
(d)
SPACE CHARGE DENSITIES
$+Q_M$
$+qP(x)$
x
$-qN_{AA}$
$-qN(x)$
(e)
(f)

Fig.413.1 Illustrating the steps of constructing the energy band diagram of a metal-oxide-semiconductor capacitor sandwich. (a) The Al metal, the SiO_2 insulator and the p-Si semiconductor are isolated and not in electrical nor physical contact but are in thermal equilibrium with a heat sink such as the ambient gas molecules. (b) The three pieces are put in physical contact with the Al and the p-Si separated by the SiO_2 which is assumed to be an ideal insulator to prevent any charge or electron transfer between the Al and the p-Si. (c) The instance of time when an electron pathway is established between the Al and p-Si by interconnecting them with a conductor wire, such as Cu, and electrons begin to transfer from the Al to the p-Si. (d) The transient current ceases after electrical equilibrium is reached via electron transfer from the Al to the p-Si which charges up the Al more positively relative to the p-Si. This gives rise to the space charge layers and the bending of the energy bands in the three layers although the bending is barely visible in the Al due to its high electron concentration. (e) The space charge distribution after electrical equilibrium is established. (f) Expanded view of the potential energy in a metal Debye length.

Table 413.1
Energy Band Parameters of Selected Metals and Semiconductors

	VALENCY	ELECTRON AFFINITY (eV)	ENERGY GAP@0K (eV)	VALENCE ELECTRON CONC. (10^{22} cm^{-3})	FERMI ENERGY (eV)	WORK FUNCTION IN VACUUM (eV)
METALS						
					E_F-E_C	
Cu	1			8.48	7.04	4.53-4.94 (4.579)
Ag	1			5.85	5.49	4.52-4.74 (4.684)
Au	1			5.90	5.53	5.37-5.47 (4.824)
Pt	1			6.62	5.96	5.65-5.70 (4.889)
Ni	2			18.3	11.74	5.04-5.35 (4.689)
Pd	?					5.0 (4.739)
Zn	2			13.1	9.40	4.9
Ti	2			5.66	5.38	4.33 (4.529)
Ta	2			5.52	5.29	4.00-4.80
V	2			6.32	5.78	4.18-5.25 (4.689)
Al	3			18.1	11.65	4.06-4.41 (4.679)
In	3			11.5	8.63	4.12
SEMICONDUCTORS						
					E_I-E_V'	
C	4		5.48I	70.5	28.89	5.0
Si	4	4.029	1.1205I	19.97	12.46	$4.029+E_C-E_F\pm0.08$
Ge	4	4.0	0.74I	17.78	11.53	$4.0\ +E_C-E_F$
Sn	4		0.082D	11.71	8.73	4.42
Pb	4					4.25
SiC	4		3.03I			
BN	4		7.5 I	67.74	28.12	
GaP	4	4.0	2.34I	19.75	12.37	$4.0\ +E_C-E_F$
GaAs	4	4.07	1.52D	17.71	11.50	$4.07\ +E_C-E_F$
CdS	4	4.8	2.56D	16.13	10.81	$4.8\ +E_C-E_F$
ZnO	4		3.42D			
ZnS	4		3.84D	20.10	12.52	
INSULATORS						
SiO_2		0.9→1	8-9 I			$0.9\ +E_C-E_I$
Si_3N_4			5.0			

I=Indirect, D=Direct.

In (413.1A) the free electron mass is used. The volume density of the valence electron in the metal, N_v in electron/cm^3, can be computed from the volume atomic density, atomic number and valency. The valence electron concentration and the Fermi energies are computed using published atomic density and the assumed valency. These are all listed in Table 413.1. The free electron mass assumption introduces some error. The Fermi energy is also computed for the semiconductors using (413.1A) and the values are listed in Table 413.1. They give an estimate of the position of the bottom edge of the valence band, $E_{V'}$, measured from the intrinsic Fermi level, E_I-$E_{V'}$. It enables us to make a quantitative alignment of the energy bands of the different materials.

The metal electron work functions from three sources are also collected and listed in the last column of Table 413.1. The single lower values are older measurements probably on contaminated surfaces whose surface impurity layer lowered the surface barrier height to give the lower work functions. The larger values given in a range were the latest measurements on the different surfaces of single crystals in Ultra-High-Vacuum (UHV). The single values in ( ) are computed from the work function difference, $\phi_{MS'}$, using $\phi_{MS'}$(data) $= \phi_M - \chi_{Si}$ which gives $\phi_M = \phi_{MS'}$(data) $+ \chi_{Si}$(data) $= \phi_{MS'} + 4.029$eV. The data were obtained from measurements in metal/Si Schottky barrier diodes described in chapter 5, and the precise photoelectric measurements of the electron affinity in crystalline Si made by Allen and Gobeli, $\chi_{Si}=4.029$eV.

In order to provide all the necessary data to construct the energy band diagram of a multi-layer solid state device structure, such as the MOSC and heterojunction transistors, we have also included the experimental electron affinity and energy gap of selected semiconductors and insulators in Table 413.1.

The metal electron work function is given by $W_M=q\phi_M=VL-F_M$ as illustrated in Fig.413.1(a). The SiO_2 energy band diagram is identical to that of an intrinsic Si. The only differences are: (i) the energy gap of SiO_2 is larger than Si, 8eV versus 1.2eV; (ii) the electron affinity in SiO_2 is smaller than Si, 0.9eV versus 4.02eV; and (iii) the valence band width of SiO_2 is smaller than Si, about 2eV versus 10eV. The Fermi level of a pure SiO_2, F_O, is located slightly above the midgap of SiO_2 because the holes are much heavier than electrons in SiO_2; while the Fermi level of the p-Si, F_P, is located below the Si midgap due to the presence of the acceptor impurity.

Figure 413.1(b) shows the E-x energy band diagram when the Al, SiO_2 and p-Si are put in contact, with SiO_2 the middle layer. Since SiO_2 is a very good insulator, there is little or no electron transfer between the metal and the p-Si even though some valence electrons in Al have higher energies than those in the p-Si, that is, the electron work function in the metal is smaller than that in the semiconductor. Thus, it takes less energy to remove the metal electrons than the semiconductor electrons so that some metal electrons would transfer to the semiconductor if there were a pathway.

Figure 413.1(c) shows the E-x energy band diagram at the instance when the Al and the p-Si are connected by a highly conductive pathway, such as a copper wire indicated by a heavy dark line, before electrical equilibrium is reached. Electrons have higher energies in the Al than in the p-Si so that some Al electrons are transferred to the p-Si via the copper wire in order to occupy the lowest possible energy levels of the system. This electron transfer is indicated by the arrow and is a transient phenomenon. It charges up the p-Si negatively and the Al positively until the lower energy levels in Si are occupied by the electrons transferred from Al, then the current ceases and the system has reached electrical equilibrium. This transient is completed in a very short time interval due to (i) the large number of valence electrons in the metal and the semiconductor, (ii) the small number of electrons required to fill up the unoccupied lower energy levels in the semiconductor, and (iii) the high conductivity of interconnection wire.

In principle, there is also a movement or redistribution of electrons and holes in the SiO_2 insulator film. However, the valence band of the SiO_2 insulator is completely filled by its valence electrons and the conduction band of the SiO_2 insulator is completely devoid of electrons at absolute zero. At room temperatures, thermal generation of electron-hole pairs across the SiO_2 energy gap is negligible since the SiO_2 energy gap is very large (> 8eV). This large energy gap also puts the SiO_2 conduction band energy levels so much higher above the Al and SiO_2 conduction band energy levels that there are essentially no electrons with sufficiently high energy to occupy these high-energy conduction band states in SiO_2. Thus, the SiO_2 will maintain the quasi-steady-state or quasi-equilibrium condition of an ideal insulator whose Fermi level lies essentially at the equilibrium intrinsic Fermi level position which is slightly above the midgap since the hole effective mass is larger than the electron effective mass in SiO_2.

Figure 413.1(d) shows the E-x energy band diagram after electrical equilibrium is reached and the electron transfer from Al to p-Si has ceased. The excess electrons in the p-Si and deficient electrons (or holes which are hard to illustrate in a metal) in the Al metal produce an electric field across the oxide and the corresponding space charge layer in Al and p-Si next to the two surfaces of the oxide layer. The Fermi level in the oxide insulator will now be tilted due to the electric field, making $dF_O/dx \neq 0$. But the current in the oxide is still zero because the electron concentration in the oxide, N_O, is zero so that $J_O = \mu_O N_O dF_O/dx = 0$ even $dF_O/dx \neq 0$.

The semiconductor energy band near the SiO_2/Si interface is now curved, known as **band bending**, due to the presence of the excess electrons transferred from the Al metal. The thickness of the band-bending layer is known as the **surface space-charge layer** or just **space-charge layer**. In this case, the net space charge in this semiconductor surface layer is negative due to the excess electrons transferred from the Al metal to the p-Si. Its thickness increases with the metal/semiconductor work function difference and decreases with the semiconductor majority carrier concentration. It can be estimated using the depletion approximation, (411.5), where V_S is the surface potential (measured relative to the bulk value at large x) or the total semiconductor energy band bending given by (412.15A). At electrical equilibrium and no d.c. applied voltage to the terminals or $V_G=0$, V_S is just the metal/semiconductor work function difference which is given by $V_S=\phi_{MS}=\phi_M-\phi_S=\phi_M-(\chi_S+\phi_F)$ if there is no oxide traps since V_O would be zero also. Consider a p-Si with $N_{AA}=10^{16}\text{cm}^{-3}$ and $n_i=10^{10}\text{cm}^{-3}$ at 300K, then $q\phi_F = (E_C-E_F) = E_C-E_I + (kT/q)\log_e(N_{AA}/n_i) = 0.56 + 0.02585\text{x}\log_e(10^{16}/10^{10}) = 0.918$ eV. Using the Al effective work function of 4.679eV from Table 413.1, then from (412.12E), $V_S = -\phi_{MS} = -(\phi_{Al}-\chi_{Si}-\phi_F) = -(4.679 - 4.029 - 0.918) = 0.268\text{V}$. Using (411.3A), the surface space-charge layer thickness in the depletion approximation is then given by

$$x_d = \sqrt{(2\varepsilon_s V_S/qN_{AA})}$$

$$= \sqrt{[(2\text{x}11.8\text{x}8.854\text{x}10^{-14}\text{x}0.268)/(1.602\text{x}10^{-19}\text{x}10^{16}]} = 0.187\mu\text{m}.$$

This depletion layer thickness is of macroscopic dimension, that is, it contains many Si atoms and valence electrons. Thus, the analysis is self-consistent and valid. An invalid or self-inconsistent example is given as follows.

As expected from the Gauss Theorem, the net charge enclosed inside a metal box must be zero, if the charge on the metal box is zero. Thus, there must also be a space-charge layer in the metal gate which must be positive and have exactly the same amount of charge to cancel the net negative space charge in the semiconductor surface space-charge layer. Due to the high concentration of the electrons in the metal, the metal space-charge layer at the Al/SiO$_2$ interface is extremely thin. It is of the order of the Debye length of the metal electrons. This metal Debye length can be computed if we had used the Fermi distribution for the electron density in the Poisson equation instead of the Boltzmann approximation for low electron concentrations. Without derivation, which can be carried out by expanding the Fermi energy of the metal as a function of the electron concentration, given by (413.1), the metal Debye length that characterizes a charge or potential fluctuation in the metal is given by

$$L_{DB} = \sqrt{2\varepsilon(E_F-E_C)/3q^2N} \qquad (413.2)$$

$$= (2\pi^2)^{1/3}\cdot(\hbar/q)\cdot(\varepsilon/3m)^{1/2}\cdot N^{-1/6} \qquad (413.2A)$$

$$= (2.372\text{x}10^{21}/N)^{1/6}\ (\text{A}) \qquad (413.2B)$$

where the metal Debye length is in Angstrom (10^{-8}cm) and the electron concentration is in number/cm^3. Using $N=1.80\times10^{23}$cm^{-3} for Al from Table 413.1, the electron Debye length in Al is 0.486A. This is much less than the lattice constant or the average spacing of the Al valence electrons which is given by $N^{-1/3}=(1.8\times10^{23})^{-1/3}=1.768$A. Thus, the macroscopic Poisson and Shockley equations are no longer valid to describe the metal space-charge layer. However, the deficiency of electrons in the Al can still be pictorially illustrated by a slight bending upwards of the Al conduction band edge as indicated in Fig.413.1(d). The resultant positive space-charge in the metal is customarily illustrated as a positive delta function (spatial-impulse function), Q_M, in the space-charge distribution diagram shown in Fig.413.1(e).

An application of the equilibrium energy band diagram is given in Figs.413.2(a) to (c) which illustrate the effect of the metal/semiconductor work function difference, $\phi_{MS} = \phi_M - \phi_S = \phi_M - (\chi_S + \phi_F)$, on the energy band bending at zero applied gate voltage. Oxide and interface traps are assumed absent. Figure (a) shows the flat-band condition or zero surface potential when there is no work function difference or $\phi_{MS}=0$. Figure (b) (for $\phi_{MS}<0$) shows positive surface potential or positive shift of the bulk semiconductor energy band relative to the metal or negative shift the metal energy band relative to the semiconductor energy band for negative ϕ_{MS}. Because $\phi_{MS}<0$, or the metal work function is smaller than the semiconductor work function, some higher energy metal electrons are transferred to the semiconductor. Figure (c) (for $\phi_{MS}>0$) shows negative semiconductor surface potential or negative shift of the bulk semiconductor energy band, i.e., positive shift of the metal energy band for positive ϕ_{MS}. Because the metal work function is larger than the semiconductor work function in this case, some semiconductor electrons are transferred to the metal. The effect of a distribution of oxide charge can be readily included. For a sheet of oxide charge at $x=x_{OT}$ from the oxide/semiconductor interface, the oxide potential line has an abrupt change of slope at x_{OT}. Positive oxide charge, Q_{OT}, is equivalent to a negative ϕ_{MS} as indicated by the definition the flat-band voltage, (410.32) or (412.15B).

A second application of the energy band diagrams of a MOSC is illustrated in Figs.413.3(a)-(e). They show the effects of the applied voltage on energy band diagram of a p-Si MOSC which has $\phi_{MS}<0$ and $V_{FB}<0$. The equilibrium energy band diagram of the MOSC on p-Si given previously in Figs.413.1(d) and 413.2(b) (the latter for n-Si) is repeated in Fig.413.3(c). The energy band diagram at flat-band, $V_G = V_{FB}$, is shown in Fig.413.3(b); in strong accumulation of majority carriers, $V_G < V_{FB} < 0$, in Fig.413.3(a); at intrinsic surface, $V_G = V_{GI} > 0$ which gives $V_S = V_F$, in Fig.413.3(d); at the threshold voltage, $V_G = V_{TH}$ and $V_S = 2V_F$, in Fig.413.3(e); and in strong inversion, $V_G > V_{TH} > 0$ and $V_S > 2V_F + 2kT/q > 0$ in Fig.413.3(f). A highly conductive or strong inversion surface channel is induced when $V_G > V_{GT}$. In integrated circuits, MOS transistors operate in this d.c. applied gate voltage range which will be discussed in chapter 6.

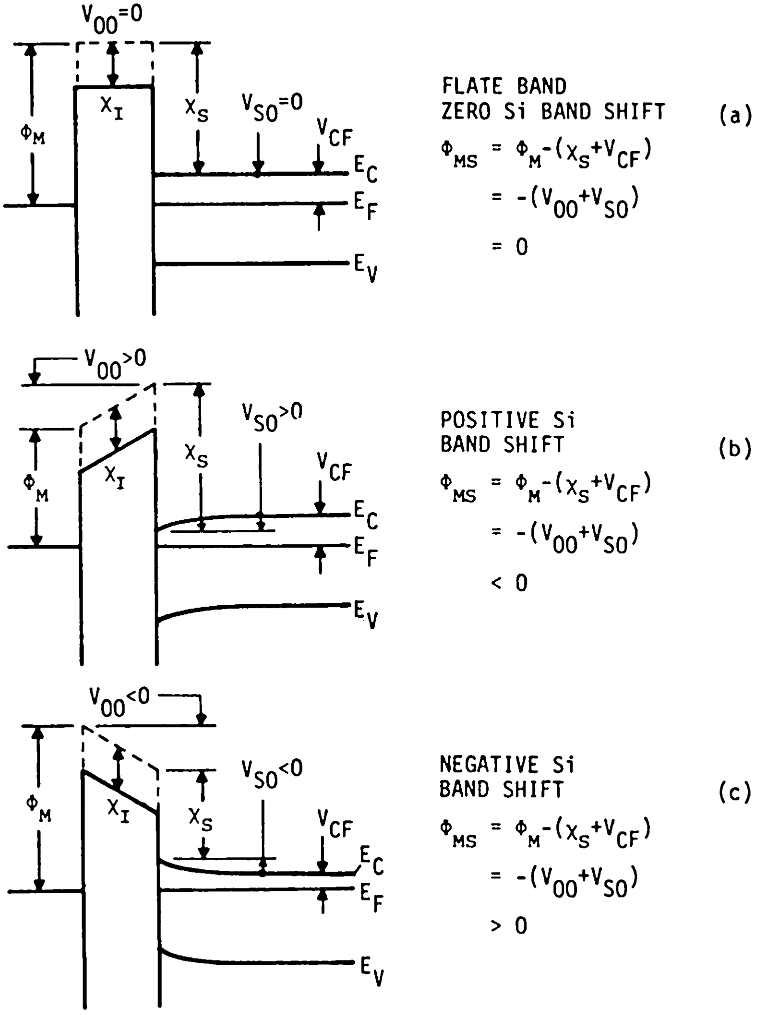

$V_{00}=0$
Φ_M
X_I
X_S
$V_{SO}=0$
V_{CF}
E_C
E_F
E_V
FLATE BAND
ZERO Si BAND SHIFT
(a)
$\Phi_{MS} = \Phi_M-(X_S+V_{CF})$
$= -(V_{00}+V_{SO})$
$= 0$
$V_{00}>0$
Φ_M
X_I
X_S
$V_{SO}>0$
V_{CF}
E_C
E_F
E_V
POSITIVE Si
BAND SHIFT
(b)
$\Phi_{MS} = \Phi_M-(X_S+V_{CF})$
$= -(V_{00}+V_{SO})$
< 0
$V_{00}<0$
Φ_M
X_I
X_S
$V_{SO}<0$
V_{CF}
E_C
E_F
E_V
NEGATIVE Si
BAND SHIFT
(c)
$\Phi_{MS} = \Phi_M-(X_S+V_{CF})$
$= -(V_{00}+V_{SO})$
> 0

Fig.413.2 Equilibrium energy band of three MOSC's showing the effect of metal/semiconductor work function difference, Φ_{MS}, on the equilibrium energy band bending. Oxide and interface traps are neglected. (a) $\Phi_{MS}=0$, flat-band. (b) $\Phi_{MS}<0$ and negative shift of the metal energy band or positive shift of the semiconductor bulk energy band. (c) $\Phi_{MS}>0$ and positive shift of the metal energy band or negative shift of the semiconductor bulk energy band.

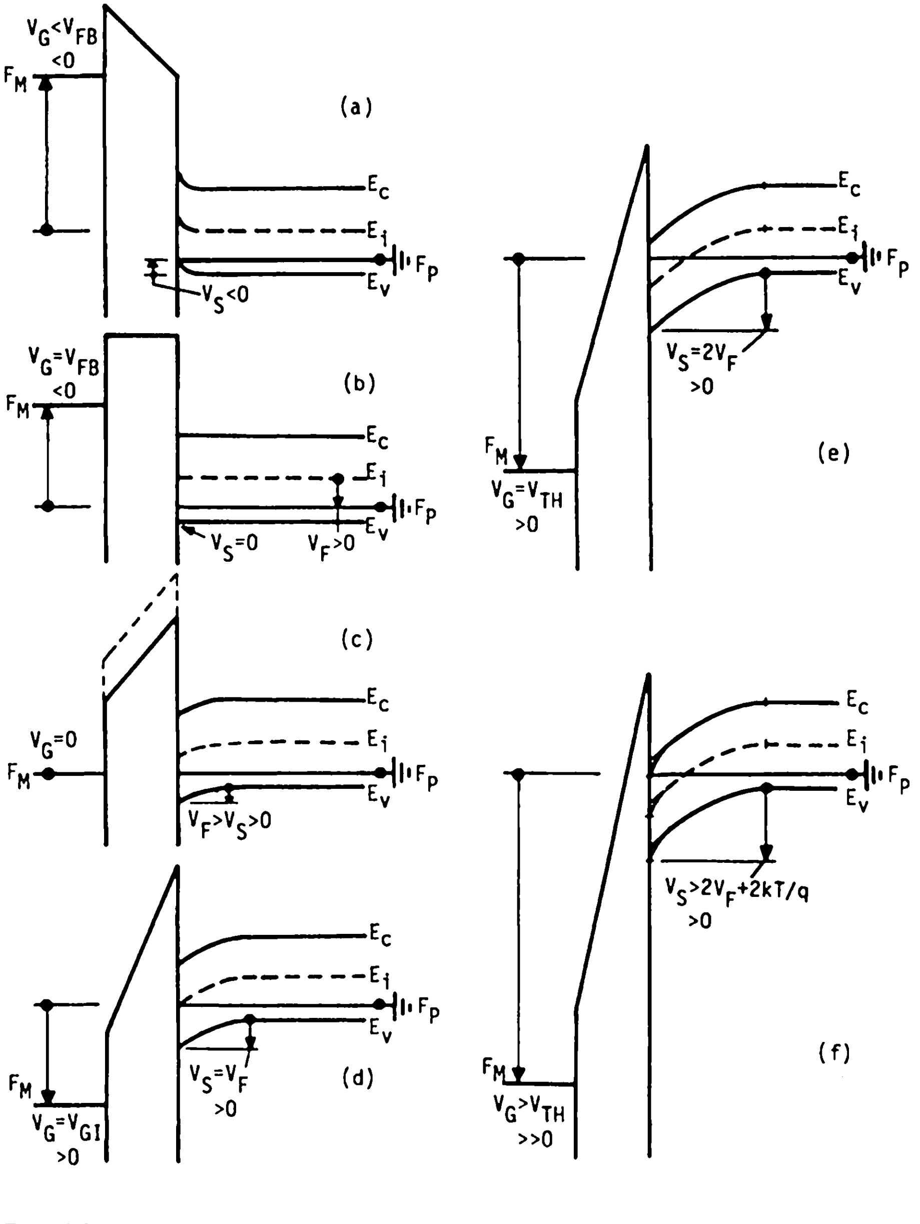

$V_G<V_{FB}$
<0
F_M
(a)
E_c
E_i
F_p
E_v
$V_S<0$
$V_G=V_{FB}$
<0
F_M
(b)
E_c
E_i
F_p
E_v
$V_S=0$
$V_F>0$
(c)
$V_G=0$
F_M
E_c
E_i
F_p
E_v
$V_F>V_S>0$
E_c
E_i
F_p
E_v
F_M
$V_S=V_F$
>0
(d)
$V_G=V_{GI}$
>0
E_c
E_i
F_p
E_v
$V_S=2V_F$
>0
F_M
(e)
$V_G=V_{TH}$
>0
E_c
E_i
F_p
E_v
$V_S>2V_F+2kT/q$
>0
(f)
F_M
$V_G>V_{TH}$
>>0

Fig.413.3 The effect of applied d.c. gate voltage on the energy band diagram of a p-Si MOSC with $\phi_{MS}<0$. (a) $V_G<V_{FB}<0$. (b) $V_G=V_{FB}<0$. (c) $V_G=0$. (d) $V_G=V_{GI}>0$. (e) $V_G=V_{TH}$. (f) $V_G>V_{TH}>0$.

420 TRANSIENTS IN MOSC

Time dependence or instability of the Si MOSCV curves was observed when the first MOSC was made in the laboratories during 1959-1961. The origin of the drift was soon identified as the migration or drift of positively charged ions, especially sodium ions Na^+ in the oxide [400.5], when an electric field or gate voltage was applied. Other origins were discovered later, including the migration of proton or hydrogen ion in the oxide, the capture of injected electrons and holes by the traps in the oxide at high electric fields, the generation of oxide and interface traps from breaking the weak Si-O and Si-Si bonds by energetic or hot electron impact and by thermal hole capture. The drift of the MOSCV curves at high electric fields was found to depend on the polarity of the applied gate voltage or the direction of the electric field because the polarity determines the directions of ion drift and electron and hole injection. This polarity dependence gives a hysteresis loop in the CV curve. The time dependence of the MOSCV curve is undesirable because it causes transistor instability. However, it has been used as a very sensitive detector in manufacturing to determine the origins that cause the MOSCV instabilities, and in research to identify and characterize the migrating species, such as the sodium ion in the SiO_2.

Soon after the identification and elimination of the oxide ion drift instability, other transient producing mechanisms become observable. In particular, the electronic mechanisms were observed which included generation-recombination-trapping-tunneling of electrons and holes at the traps at the SiO_2/Si interface, and in the surface space-charge layer and bulk of Si. Once their origin was understood, they were used to detect and characterize the kinetic properties of the electronic traps at the SiO_2/Si interface and in the Si surface space charge layer. The six parameters of interface and bulk traps described in chapter 3 were characterized which were c_n, c_p, e_n, e_p, E_T, and N_{TT}. Differential or derivative methods were developed to measure the capture and emission rate coefficients as a function of the electric field and the trap concentration as a function of position. Processes listed in Table 360.1 which were investigated by the MOSC transients included the thermal, optical and impact energy exchange mechanisms. The theory of the transients and measurement methodologies and experimental data for gold-doped Si MOSC's to illustrate the methodologies were presented in 1972 [420.1, 420.2].

[420.1] C.T.Sah and H.S.Fu, "Current and capacitance transient responses of MOS capacitor, I. General theory and applications to initially depleted surface without surface states," physica status solidi (a)**11**, pp.297-310, 16 May 1972.

[420.2] C.T.Sah and H.S.Fu, "Current and capacitance transient responses of MOS capacitor," II. Recombination centers in the surface space-charge layer," physica status solidi (a)**14**, pp.59-70, 16 November 1972.

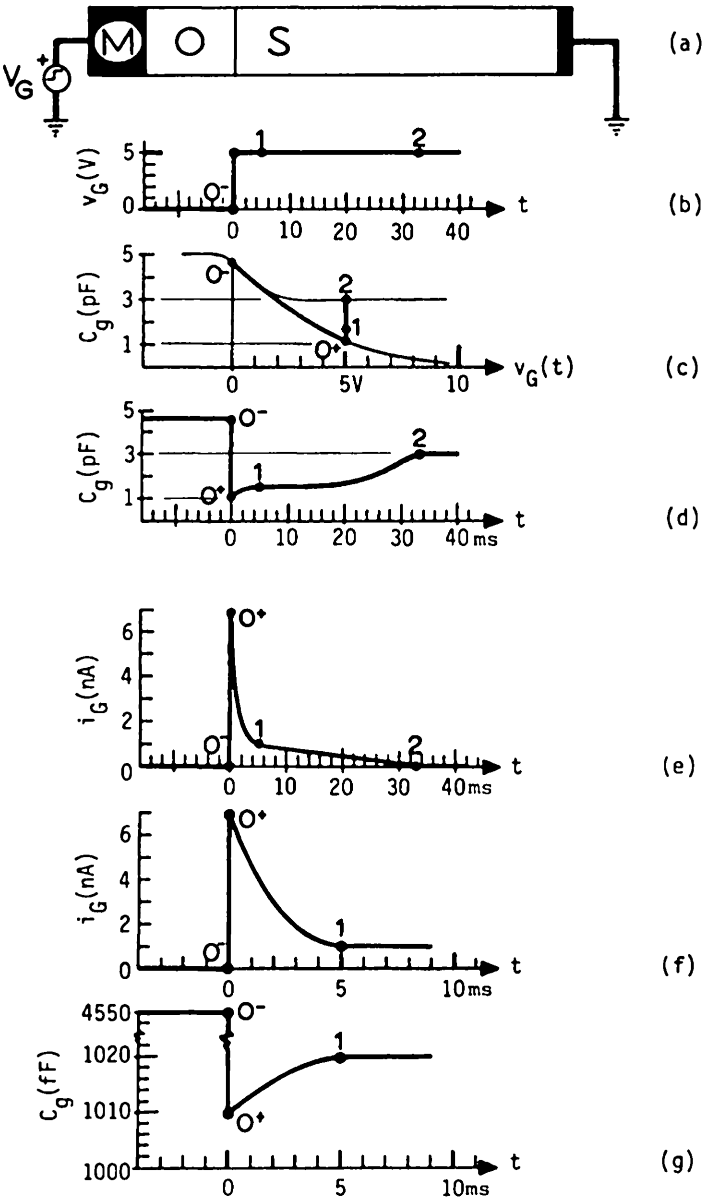


Fig.420.1 High-frequency capacitance and d.c. current transients in a p-Si MOSC. (a) Cross-sectional view. (b) Applied voltage step. (c) Locus of $C_g(t)$ vs $v_G(t)$. (d) $C_g(t)$ vs t. (e) Phase I of $C_g(t)$. (f) $i_G(t)$ vs t. (g) Phase I of $i_G(t)$.

The high-frequency (1-MHz) capacitance transient, $C_g(t)$, and its corresponding d.c. current transient, $i_G(t)$, are illustrated by the p-Si MOSC example shown in Figs.420.1(a)-(e). As shown in Fig.420.1(a), the gate voltage is switched from zero $v_G(t=0^+)=0$ [or from a negative or forward gate bias in strong accumulation $v_G(0^-)=-V_F<<0$] to a large reverse bias, $v_G(t=0)=V_R=+5V$. The locus of the capacitance on the C-V plane is shown in Fig.420.1(c). The capacitance vs time or the capacitance waveform is shown in Fig.420.1(d). These time dependences are explained in the following paragraphs.

The C_g-v_G locus in Fig.420.1(c) drops abruptly from C_{0-} before the voltage step to C_I, the depletion value at $v_G=+5V$. The locus follows the depletion C-V curve when the gate voltage is abruptly stepped from 0V to +5V because negligible minority carriers (electrons) are generated during the very fast voltage step. Afterwards, C will gradually rise from C_I towards the final state-steady high-frequency value in strong inversion as the minority carriers (electrons) are attracted to or collected at the SiO_2/Si interface by the positive v_G. **The minority carrier accumulation** will then **invert** the gated p-Si surface to n-type when $n(x=0,t) > p(x=0,t)$.

Figure 420.1(d) shows that there are two phases in the capacitance transient: a small and fast phase I which is enlarged in Fig.420.1(e), and a slow and large phase II. There is also a corresponding d.c. gate current transient due to the gate voltage step. The current transient is shown in Fig.420.1(f) and its phase I is expanded in (g). In contrast to the capacitance transient, the phase I current change is fast and large while the phase II, slow and small. **Phase I** is purely exponentially dependent on time and is due to the emission (by thermal, optical or impact) of the trapped holes from the interface and space-charge-layer traps which is followed by emission of trapped electrons until the steady-state trapped charge distribution is reached. Thus, phase I will be known as **detrapping phase. Phase II** is non-exponential and it is due to the **accumulation** of the generated or trap-emitted minority carriers (electrons) at the SiO_2/Si interface which is eventually **inverted** to n-type at the end of phase II when $C=C_F$. Thus, phase II will be known as the **inversion** phase.

421 Capacitance Transients

The depletion capacitance formula, (411.8A), can be used to predict the two-phase capacitance transients just described without doing any algebra. (Again, the basic physics, choice of notation, and appearance of the formulae have eliminated the tedious algebra which is undertaken in the next section, 421, to provide further insight using the E-x energy band diagram.) The contributions to the capacitance transient from the interface and bulk hole traps can be readily included in and then extracted from (411.8A). During phase I, the interface hole trap will give a time-dependent interface charge which is included in $q_{IT}(t)$ in the V_{FB} term. The bulk hole trap will give a time-dependent bulk charge which can be included in N_{AA} by replacing it with $N_{AA}-p_T(t)$. The negative sign appears since N_{AA} is negatively charged while the occupied hole trap, p_T, is positively charged. During phase II, the accumulation of the minority carriers (electrons) at the interface can also be included by adding $q_N(t)$ to the V_{FB} term. Thus, (411.8A) becomes

$$\frac{1}{C_g^2} = \frac{1}{C_o^2} + \frac{2\{(V_G - V_{FB}) + [q_{IT}(t) - |q_N(t)|]/C_o\}}{\varepsilon_s q[N_{AA} - p_T(t)]}. \tag{421.1}$$

For phase I, $q_N=0$, and the capacitance change is very small because the terms q_{IT} and p_T are very small. Thus, the above can be expanded to give

$$\begin{aligned} -2\delta C_g(t)/C_g(0^+)^3 &= + 2\Delta q_{IT}(t)/(C_o\varepsilon_s qN_{AA}) \\ &\quad + 2[(V_G - V_{FB})/\varepsilon_s qN_{AA}][\Delta p_T(t)/N_{AA}] \end{aligned} \tag{421.2A}$$

or

$$\begin{aligned} \delta C_g(t)/C_I^3 &= + (x_o/\varepsilon_o\varepsilon_s)[q_{IT}(0)-q_{IT}(t)]/(qN_{AA}) \\ &\quad + (C_I^{-2}-C_o^{-2})[p_T(0)-p_T(t)]/2N_{AA}. \end{aligned} \tag{421.2B}$$

The definition $C_g(0^+)=C_I$ and the depletion capacitance-voltage relationship from (411.8A), $C_I^{-2}-C_o^{-2}=2(V_G-V_{FB})/(\varepsilon_s qN_{AA})$, are used in (421.2A) to give (421.2B).

The result given in (421.2B) shows the two contributions to phase I: the first term comes from interface traps, and the second term, the bulk traps in the surface space-charge layer. Both are from hole detrapping which adds negative space charge to the negatively charged acceptor, N_{AA}, thus, both cause an increase of capacitance with time as observed in the experiment. To obtain an explicit time dependence, the sixth Shockley equation, (350.6), must be solved to give the time dependences of the charge trapped at the interface and bulk traps. Consider the thermal (SRH) and extrinsic optical mechanisms given by (372.1), and use the identity $n_T(x,t)+p_T(x,t)=N_{TT}(x)$, then Shockley's sixth equation (350.6) becomes

$$+\ \partial n_T/\partial t = (g_P-r_P) - (g_N-r_N) \tag{350.6}$$

$$\equiv\ -\ \partial p_T/\partial t = (e_p p_T - c_p p n_T) - (e_n n_T - c_n n p_T) \tag{421.3}$$

$$= (e_p+c_p p+e_n+c_n n)p_T - (c_p p+e_n)N_{TT} \tag{421.4}$$

$$\simeq (e_p+e_n)p_T - e_n N_{TT} \tag{421.4A}$$

where $c_n=c_n^t+c_n^o$, $e_n=e_n^t+e_n^o$, etc. by combining the thermal and optical band-trap transitions into one term. The depletion approximation, $p=0$ and $n=0$, is used in (421.4) to give (421.4A). Hole depletion, $p=0$, is valid, because during phase I, the large electric field in the Si surface space charge layer sweeps out all the majority carriers (holes) from the interface and the surface space-charge layer into the bulk. Electron depletion, $n=0$, is valid in the Si surface space-charge layer also because of the large electric field. However, electrons are swept to the oxide/Si interface. But phase I is so short that few electrons are accumulated so that $q_N(x=0,t<t_1)\approx 0$. The solution of (421.4A) can be written down immediately if we assume that the all the traps are filled by holes initially. This is

$$p_T(x,t) = [e_nN_{TT}/(e_p+e_n)] + [e_pN_{TT}/(e_p+e_n)]\exp(-t/\tau_1) \qquad (421.5)$$

where

$$\tau_1 = 1/(e_p+e_n) \qquad (421.6)$$

is the detrapping time constant. This is expected: the detrapping rate is the sum of hole and electron detrapping rates, e_p+e_n. Thus, for the interfacial surface states, we let $N_{TT}=N_{SS}\delta(x)$ and use subscript s for interfacial surface state, then,

$$\begin{aligned} q_{IT}(0) - q_{IT}(t) &= \int_{0^-}^{0^+} q[p_{IT}(x,0) - p_{IT}(x,t)]dx \\ &= \int_{0^-}^{0^+} q[e_{ps}/(e_{ps}+e_{ns})][1 - \exp(-t/\tau_1)]N_{SS}\delta(x)dx \\ &= + q[e_{ps}/(e_{ps}+e_{ns})][1 - \exp(-t/\tau_1)]N_{SS}. \qquad (421.7) \end{aligned}$$

The phase I capacitance transient due to detrapping at the interface traps is obtained using this result in the first term of (421.2B). It is given by

$$\begin{aligned} \delta C_g(t)/C_I^3 &= + (x_o/\varepsilon_o\varepsilon_s)[q_{IT}(0)-q_{IT}(t)]/(qN_{AA}) \\ &= + (x_o/\varepsilon_o\varepsilon_s)(N_{SS}/N_{AA})[e_{ps}/(e_{ps}+e_{ns})][1 - \exp(-t/\tau_{1s})]. \qquad (421.8) \end{aligned}$$

where $\tau_{1s}=(e_{ps}+e_{ns})^{-1}$. This result shows that experimental measurements can be made to determine the thermal and optical properties of the interface traps, e_{ps} and e_{ns}, and interface trap density, N_{SS}. The physics of thermal capture and emission discussed in section 381 shows that the thermal emission rate is given by (381.4) and thermal capture rate is a product of the capture cross section and the thermal velocity. Applying to the interfacial surface state, then,

$$\begin{aligned} e_{ps}{}^t &= c_{ps}{}^tN_V\exp[-(E_{TS}-E_V)/kT] \qquad (381.4) \\ &= \sigma_{ps}{}^t\theta_pN_V\exp[-(E_{TS}-E_V)/kT] \qquad (421.9) \end{aligned}$$

and

$$e_{ns}{}^t = \sigma_{ns}{}^t\theta_nN_C\exp[-(E_C-E_{TS})/kT] \qquad (421.10)$$

These relationships show that if the phase I time constant and total capacitance transient are measured as a function of temperature, then the slope of the Arrhenius plot, τ_{1s} will give the energy level of the interface trap relative to the conduction and valence band edges.

A similar derivation gives the phase I capacitance transient due to hole detrapping at the bulk traps in the Si surface space-charge layer. This is obtained using (421.5) for $p_T(0)$-$p_T(t)$ in the second term of (421.2B) to give

$$\delta C_g(t)/[C_I{}^3(C_I{}^{-2}-C_o{}^{-2})]$$

$$= (N_{TT}/2N_{AA})[e_{pb}/(e_{pb}+e_{nb})][1 - \exp(-t/\tau_{1b})]. \qquad (421.11)$$

As an numerical example using gold acceptor in Si shown in Fig.381.2, $e_{nb}=1000\text{s}^{-1}$ and $e_{pb}=100\text{s}^{-1}$ at $T=300\text{K}$ so that $\tau_{1b}=1/1100=9.9\text{ms}$. Lower the temperature to 200K, $e_{nb}=10^{-2}\text{s}^{-1}$ and $e_{pb}=2\text{x}10^{-4}\text{s}^{-1}$ so that $\tau_{1b}=1/0.0102=98\text{s}$. Thus, accurate capacitance transient curve can be measured by lower the temperature to slow down the response.

The phase II capacitance transient formulae can be derived also from the general result, (421.1), by noting that the transient is entirely due to minority carrier (electron) accumulation at the oxide/Si interface. Thus, we retain only the term $q_N(t)$ and (421.1) becomes

$$C_g{}^{-2} = C_o{}^{-2} + 2[(V_G-V_{FB}) - |q_N(t)/C_o|]/(\varepsilon_s qN_{AA}) \qquad (421.12)$$

$$= C_I{}^{-2} - 2(x_o/\varepsilon_o\varepsilon_s)[|q_N(t)|/(qN_{AA})]. \qquad (421.13)$$

$$= C_I{}^{-2} - 2C_o{}^{-2}(\varepsilon_o/\varepsilon_s)[|q_N(t)|/(qN_{AA}x_o)]. \qquad (421.13\text{A})$$

The inversion charge, $q_N(t)$, is entirely due to the accumulation of the electrons emitted from the interface and bulk traps. It is given by

$$\partial q_N/\partial t = \int_{0^-}^{\infty} qe_n n_T(x)dx$$

$$= q[e_{ns}e_{ps}/(e_{ps}+e_{ns})]N_{SS} + q[e_{nb}e_{pb}/(e_{nb}+e_{pb})]N_{TT}x_d(t) \qquad (421.14)$$

and

$$q_N(t) = q[e_{ns}e_{ps}/(e_{ps}+e_{ns})]N_{SS}t$$

$$+ q\int_0^t [e_{nb}e_{pb}/(e_{nb}+e_{pb})]N_{TT}x_d(t)dt. \qquad (421.15)$$

The result shows that the phase-II capacitance transient is non-exponential because the change is large. The interfacial trap gives a simple time dependence, i.e. $C_g{}^{-2}$ is proportional to t. Using (421.15) in (421.13) it is

$$C_g{}^{-2} = C_I{}^{-2} + 2(x_o/\varepsilon_o\varepsilon_s)(N_{SS}/N_{AA})[e_{ns}e_{ps}/(e_{ns}+e_{ps})]t. \qquad (421.16)$$

The time dependence of $C_g(t)$ due to bulk trap is more complex. If the bulk trap dominates, then the integral in (421.15) is the only term and (421.13A) can be solved analytically using $x_d(t)/\varepsilon_s = C_d^{-1} = C_o^{-1} - C_g^{-1}$. The solution is

$$(N_{TT}/N_{AA})[e_{nb}e_{pb}/(e_{nb}+e_{pb})]t = -\int_{C_I}^{C_g} \frac{(C_o/C)d(C_o/C)}{(C_o/C) - 1} \qquad (421.17)$$

$$= (C_o/C_I) - (C_o/C_g) + \log_e\{[(C_o/C_I)-1]/[(C_o/C_g)-1]\} \qquad (421.18)$$

Thus, the phase II duration is given by

$$(N_{TT}/N_{AA})[e_{nb}e_{pb}/(e_{nb}+e_{pb})]t_2 = k ,$$

$$= (C_o/C_I)-(C_o/C_F) + \log_e\{[(C_o/C_I)-1]/[(C_o/C_F)-1]\} \qquad (421.19)$$

For $C_I = 0.25C_o$ and $C_F = 0.5C_o$, $k=3$. Thus, the phase II duration is a few times $[(e_{nb}+e_{pb})/e_{nb}e_{pb}](N_{AA}/N_{TT})$. This shows that it is amplified by the ratio N_{AA}/N_{TT} which can be more than 1000 if the bulk trap concentration is very low. Physically, the ratio comes about because it takes about N_{AA} electrons to invert the surface which has N_{AA} holes initially but electrons are generated by only N_{TT} bulk traps. Again using gold acceptor in Si as the example and taking a typical VLSI value of $N_{AA}/N_{TT} = 10^{15}/10^{11} = 10^4$, then the phase II inversion time is $t_2 = 3[(1000+100)/1000\text{x}100]\text{x}10^4 = 330\text{s} = 5.5$minutes at 300K. This is longer than normally observed due to contribution from interface traps. For example, if $x_o = 1000\text{A} = 10^{-5}\text{cm}$ and $N_{SS} = 10^9\text{cm}^{-2}$, then (421.16) shows that a gold-like interface trap would give a faster phase II transient at 300K with a duration of $t_2 = (4^2-2^2)(12/4)(1/2)(10_{15}\text{x}10^{-5}/10^9)(1000+100)/(1000\text{x}100) = 1.98\text{s}$. It is 165 times faster than the bulk traps.

422 Current Transients

Current transients in MOSC arise from the same origin as the capacitance transients stated in section 420. A more detailed formulation is desirable to analyze the current transient waveform because it contains both the capacitance (or displacement) current and a conduction current due to majority carriers (holes in the p-Si MOSC example) flowing into the Si substrate and the contact on the back surface of the Si wafer. The current transients can be analyzed via the charge control method without using the E-x energy band diagram, like the capacitance transient analyses given in section 421. However, we shall take this opportunity to use the E-x energy band diagram because it helps to gain additional physical insights and it provides a visual model on the mechanisms.

The energy band diagrams of a p-Si MOSC during the phase I and II transients are shown in Figs.422.1(a)-(e) and 422.2(f)-(h) for a p-Si MOSC shielded from light. These are described in the following paragraphs with the charge control analyses.

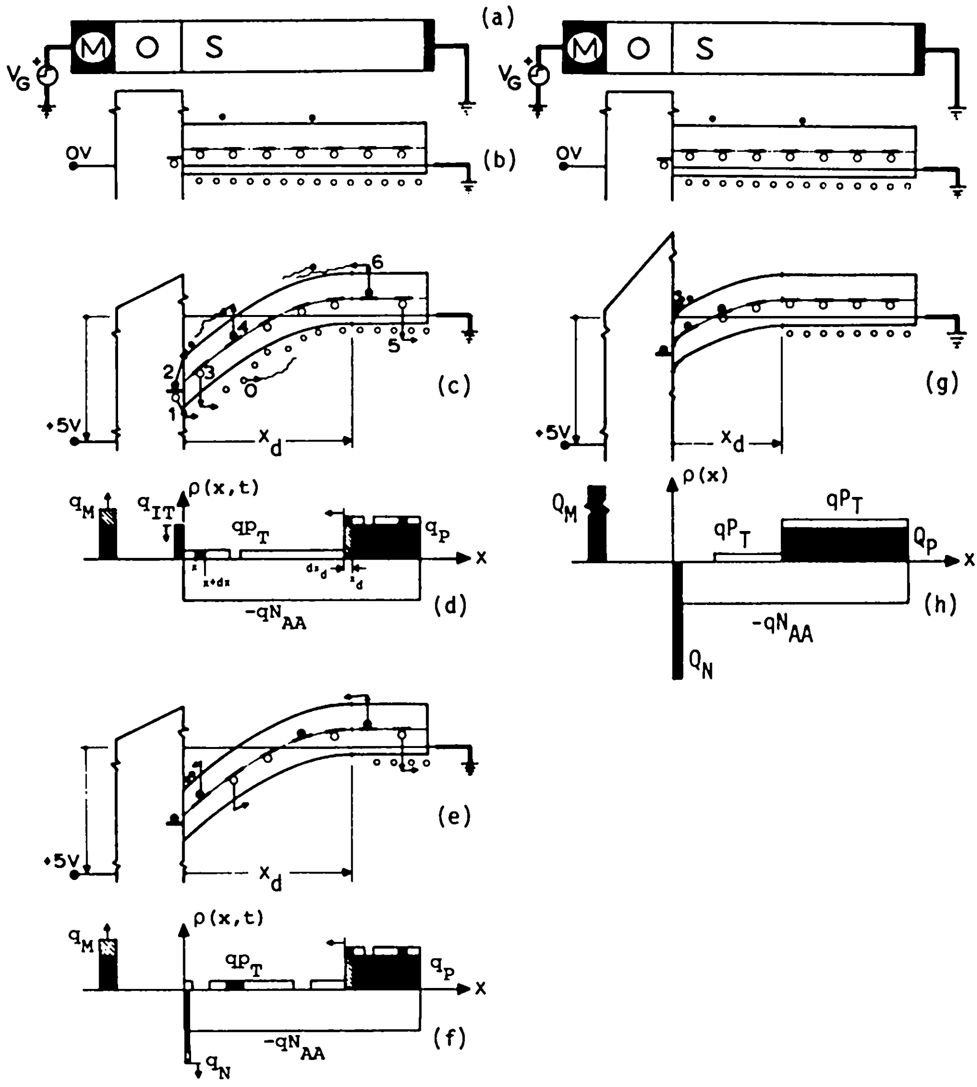

(a)
M
O
S
V_G
+
(b)
0V
(c)
1
2
3
4
5
6
+5V
x_d
(d)
q_M
q_{IT}
$\rho(x,t)$
qp_T
q_P
x
x+dx
dx_d
x_d
$-qN_{AA}$
(e)
+5V
x_d
(f)
q_M
$\rho(x,t)$
qp_T
q_P
x
$-qN_{AA}$
q_N
(g)
+5V
x_d
(h)
Q_M
$\rho(x)$
qP_T
qP_T
Q_P
x
$-qN_{AA}$
Q_N

Fig.422.1 The dark energy band and space-charge diagrams of the p-Si MOSC during a depletion-inversion voltage step transient. (a) Cross-sectional view. (b) Prior to voltage step. (c) and (d) Immediately after voltage step. (e) and (f) Inversion started $t>t_1$. (g) and (h) Steady-state inversion reached, $t>>t_2$.

Before the voltage step, we assume a flat band shown in Fig.422.1(b) which illustrates that all the interface and bulk traps are occupied by holes. When the gate voltage is switched from V_{GFB} or a more forward (negative) value to +5V or a large (positive) reverse value, holes in the surface space charge layer are swept out into the bulk in a drift transit time, $t=x_d/\theta_{psat}=10^{-5}cm/10^{7}(cm/s)=10^{-12}s$. This gives a large and very fast current spike which is normally not observed since it is too fast and the current pulse is too short. After this initial sweep-out, labeled 0 in Fig.422.1(c), trapped holes are emitted from the interface traps and from the bulk traps in the surface space-charge layer into the Si valence band. These transitions are labeled 1 and 3 in Fig.422.1(c). These released or emitter holes are instantaneously swept out of the space-charge layer into the Si bulk in about 1ps by the high electric field. Thus, detrapping holes is rate limited by the thermal emission of holes from the traps if light is absent, or by optical emission if light is present. When trapped holes are emitted, the left behind trapped electrons can then be emitted to the conduction band which are labeled by 2 and 4 in Fig.422.1(c). These electrons are then swept to the oxide/Si interface by the high electric field. Some are captured by the interface traps and the remaining are accumulated at the interface. The two hole currents (one from interface and one from bulk traps) contribute the conduction component of the gate current. Since they are flowing into two parallel capacitance, C_o and C_{xd}, as indicated by the equivalent circuit Fig.422.1(f), the fraction that flows in C_o which gives part of the gate current, $j_G(t)$, is given by

$$j_G(t)_{hole} = [C_o/(C_o+C_{xd})]q[e_{ps}p_{TS}(t) + e_{pb}p_{TB}(t)x_d(t)] \qquad (422.1)$$

The displacement current due to the variation with time of the trapped charge located between x and d+dx, $p_{TB}(x,t)dx$, is also divided between two parallel capacitance branches shown in Fig.422.1(f). This can be computed as follows. The charge induced by $p_{TB}(x,t)$ is divided by the two capacitances $C_x=\varepsilon_s/x$ and $C_{xd-x}=\varepsilon_s/(x_d-x)$. Thus, the charge fluctuation induced at x_d is

$$[C_{xd-x}/(C_{xd-x}+C_x)]q\partial p_{TB}(x,t)/\partial t = (x/x_d)q\partial p_{TB}(x,t)/\partial t \qquad (422.2)$$

and the total charge variation induced at x_d due to all the time-dependent trapped charges in the space-charge layer is obtained by integrating the above over the space charge layer. This is an effective current source driving the oxide/Si interface or the two capacitances in parallel, C_o and C_{xd}. Thus, its component flowing through the gate is reduced by the ratio $C_o/(C_o+C_{xd})$, which is

$$j_G(t)_{displacement} = [C_o/(C_o+C_{xd})]\int_{0+}^{x_d} q(x/x_d)[\partial p_{TB}(x,t)/\partial t]dx$$

$$= [C_o/(C_o+C_{xd})]q(x_d/2)\partial p_{TB}(t)/\partial t \qquad (422.3)$$

The total gate current is the sum of these two components, the hole conduction current from (422.1) and the displacement current from (422.3). Using the trapping rate equation (421.4A) for the depletion layer, then the total current is

$$
\begin{aligned}
j_G(t) &= [C_o/(C_o+C_{xd})]q\cdot\Big[e_{ps}p_{TS}(t) + ½[e_{pb}p_{TB}(t)+e_{nb}n_{TB}(t)]x_d(t)\Big] \\
&= [C_o/(C_o+C_{xd})]q\cdot \\
&\quad \Big[+ N_{SS}\{[e_{ps}e_{ns}/(e_{ps}+e_{ns})] + [e_{ps}{}^2/(e_{ps}+e_{ns})]\exp(-t/\tau_{1s})\} \\
&\quad + N_{TT}x_d\{[e_{pb}e_{nb}/(e_{pn}+e_{nb})] \cdot [1+½[(e_{pb}/e_{nb})-1]\exp(-t/\tau_{1b})\}\Big]
\end{aligned}
\tag{422.4}
$$

The above general result gives both the phase I and II current transients. It may be derived more rigorously using the Shockley equations. Various forms of the results were given in [420.1] and [420.2]. Note that two important asymptotic solutions can be used to check the theoretical derivation. If the oxide is very thick, then C_o is very small and the gate current is proportional to the oxide thickness. If the oxide is very thin, then C_o is very large and the gate current is independent of the oxide thickness. The second result is exactly that obtained from a p/n junction space-charge layer to be described in chapter 6 which will require no further algebra.

For phase I, $x_d(t)$ is approximated by $x_d(0)$, i.e. the initial value measured by the capacitance $C_I = C_oC_{xd0}/(C_o+C_{xd0})$, and it is assumed constant since N_{SS} and N_{TT} are small. Thus, there are two exponential decay curves, one from the interface trap and one from the distributed bulk trap in the surface space-charge layer. The interface trap is assumed to be a single-level trap but in practice, there is a distribution of interface traps in the energy gap and in space going into the oxide with very different electron and hole emission rates. Superposing the many exponentials would give a highly nonexpoential decay. If the interface trap is negligible and the bulk trap dominates, then the decay is truly exponential, such as the decay observed on gold-doped Si MOSC shown in Fig.420.1(g).

For phase II, $t>3\tau_{1s}$ or $t>3\tau_{1b}$ whichever one is larger, then the densities of the trapped interface and bulk charges have reached the quasi-steady-state value, i.e. $\partial p_{TS}/\partial t=0$ and $\partial p_{TB}/\partial t=0$. The gate current transient is then given by

$$
\begin{aligned}
j_G(t) &= [C_o/(C_o+C_{xd})]q\cdot \\
&\qquad \Big[+ N_{SS}\{[e_{ps}e_{ns}/(e_{ps}+e_{ns})] + N_{TT}x_d\{[e_{pb}e_{nb}/(e_{pn}+e_{nb})]\Big]
\end{aligned}
\tag{422.5}
$$

The decay is determined by the decrease of the depletion layer thickness, $x_d(t)$, which can be obtained using $x_d(t)/\varepsilon_s = C_d^{-1} = C_o^{-1} - C_g^{-1}$ in the transcendental equation (421.16) if interface trap dominates, and (421.18) if bulk trap dominates. The decay is nearly linear with time for both cases which is consistent with the experimental data of gold bulk trap dominant in Fig.420.1(f). When interface trap dominates, (421.16) can be expanded and the linear term retained to give

$$x_d(t) \approx x_d(0) - (C_I/C_O)(N_{SS}/N_{AA})[e_{ns}e_{ps}/(e_{ns}+e_{ps})]t \tag{422.6A}$$

$$= x_d(0) - \left[\frac{\varepsilon_s x_o}{\varepsilon_s x_o + \varepsilon_o x_d(0)}\right](N_{SS}/N_{AA})[e_{ns}e_{ps}/(e_{ns}+e_{ps})]t \tag{422.6B}$$

If bulk trap dominates, (421.18) can be expanded to give

$$x_d(t) = x_d(0) - (\varepsilon_s/\varepsilon_o)x_o\{(N_{TT}/N_{AA})[e_{nb}e_{pb}/(e_{nb}+e_{pb})]t + \log_e[x_d(0)/x_d(t)]\} \tag{422.7}$$

$$\approx x_d(0) - (\varepsilon_s/\varepsilon_o)x_o(N_{TT}/N_{AA})[e_{nb}e_{pb}/(e_{nb}+e_{pb})]t \tag{422.7A}$$

The above solution for phase II terminates when $C_g(t) = C_F$ in (421.16) for interface trap dominant or in (421.18) when bulk trap dominant. This sharp cut off is due to the depletion approximation $n=p=0$ in the space-charge layer instead of $np=n_i^2$ at equilibrium which neglected the capture of electrons and holes by the traps. This does not give a large error, however.

It is evident from the phase I and II solutions of the current transients, fundamental parameters of the interface and bulk traps can be measured. For example, the phase I current transient by itself will give all six parameters of an interface or bulk trap. The reason is rather simple. Let us assume that the majority carrier emission rate is much larger than the minority carrier emission rate, i.e., $e_p >> e_n$. Then, the large initial current and its decay rate depends mainly on the faster majority carrier (hole) emission or e_p because most of the traps are initially filled by majority carriers, while the final quasi-steady-state current depends mainly on slower minority carrier (electron) emission e_n because the two emission processes are in series, i.e. one after the other, so that the slower one determines the overall rate.

The phase II current transient can also be used as additional determination of the parameters of the traps. It is essentially a measurement of the thickness of the depleted space-charge layer. Equations (422.6B) and (422.7A) show that the slope is proportional to the steady-state generation rate of electrons and holes, $N_{TT}e_{nb}e_{pb}/(e_{nb}+e_{pb})$ for the bulk trap in the Si surface space-charge layer. A similar expression is obtained for the interface trap at the oxide/Si interface.

430 EXACT SMALL-SIGNAL EQUIVALENT CIRCUIT OF MOSC

The small-signal equivalent circuits given in the preceding charge-control analyses are one-lump models whose circuit elements are **differential** quantities and not the impedances or admittances defined in the small-signal sinusoidal steady state. To obtain the exact equivalent circuit model of the MOSC for small-signal sinusoidal steady-state, which is valid at all frequencies and which includes also all the losses due to scattering and recombination-generation-trapping, one must differentiate the **differential** and **true** small-signal sinusoidal conductances and capacitances. Furthermore, since the MOSC has a finite thickness there is a signal delay due to the transit of the signal electrons and holes through the semiconductor layer. This distributed effect cannot be derived and modeled correctly using the charge-control analysis since charge-control lumps the semiconductor into one lump, hence, signal propagation delay due to diffusion and drift are excluded at the onset. The only way to get the correct solutions is to decompose the Shockley equations into a set of d.c. equations and a set of small-signal time-dependent equations, and to solve them simultaneously. The solution of the small-signal sinusoidal rms current divided by the assumed small-signal sinusoidal rms voltage then gives the input admittance of the MOSC. The real part of the input admittance is then the MOSC gate conductance, G_g, while the imaginary part of the input admittance is then the MOSC gate capacitance times the angular frequency, ωC_g. This straight forward but tedious mathematical analysis is deferred to an advanced course.

The exact d.c. and small-signal equivalent circuits of the one-dimensional MOSC or any one-dimensional semiconductor device have many more elements than the one-lump approximation used in the charge-control analysis shown in Figs.410.1(c) and (d). Nevertheless, they are rather easy to place in the exact equivalent circuits if the simple physics are remembered. The following physical reasoning is sufficient to draw the exact general equivalent circuits. (i) There are three current lines, the electron and hole conduction currents and the displacement or dielectric capacitance current. (ii) There are four voltage nodes from the three concentrations or quasi-Fermi potentials (v_P for holes, v_N for electrons, and v_T for trapped electrons), and the electric potential (v_I). (iii) There are four capacitances from the three shut charge storage capacitances (C_p for holes, C_n for electrons and C_t for trapped electrons), and the series dielectric capacitance C_j. (iv) There are give resistances representing the scattering and generation-recombination-trapping energy loss mechanisms in Δx: two series resistances $R_n=\Delta x/(q\mu_n n)$ and $R_p=\Delta x/(q\mu_o p)$ due to scattering, one shut conductance $G_{pn}\Delta x$ due to interband generation-recombination, two shut conductances $G_{pt}\Delta x$ and $G_{nt}\Delta x$ due to hole and electron capture-emission at a trap where the subscripts denote the three nodes (p, n, and t). The simplicity in describing the mechanisms reflects the systematic categorization of the basic physics made in chapters 2 and 3. The description shows that the transport mechanisms are all representable by circuit elements in the equivalent circuit model.

499 BIBLIOGRAPHY

There are few books that contain a comprehensive coverage of the characteristics of MOS capacitor. Some of the few books with MOSC chapters have presented the theory using unconventional notations and handbookish rather than pedagogical route. Thus, the list is limited to one intermediate level textbook, a review, and several references on state-of-the-art data and theory surveys.

[499.1] S. M. Sze, **Physics of Semiconductor Devices**, 2nd ed., chapter 7, "MIS diode and charge-coupled devices," pp.362-430, John Wiley & Sons, Inc. 1981.

[499.2] L. E. Katz, "Oxidation," chapter 4 in **VLSI Technology**, edited by S. M. Sze, 2nd ed., McGraw-Hill Book Company, 1988.

[499.3] I. W. Boyd, et. al., "Oxidation," chapter 16; C. T. Sah, et.al., "Insulating layers on Si substrate," chapter 17; and H. R. Philipp, "Silicon Oxides: Optical Functions," chapter 28; all in **Properties of Silicon**, INSPEC, The Institution of Electrical Engineers, London, 1988. 445 Hoes Lane, P.O.Box 1331, Piscataway, NJ 08855-1331.

499 PROBLEMS

P400.1 Why is $i=C(dV/dt)$ inconsistent with $i=dQ/dt$ if $Q=CV$ is used? Is $C=dQ/dV$ consistent with $i=C(dV/dt)$ and $i=dQ/dt$? Explain concisely, succinctly.

P401.1 The n-channel Si MOS transistor switch in the DRAM (dynamic random access) cell of the next generation 16 Mbit chip (1 bit = 1 cell which contains 1 transistor and 1 capacitor) requires an oxide of 100A or 10nm (1A = 10^{-8}cm = 0.1nm and 1nm = 10^{-9}m = 10^{-7}cm = 10A). The cell sizes or lithographic line width are about 0.7µm = 7000A = 700nm in order to pack 16 million cells on one 0.5cmx0.5cm chip. This small line width requires low temperature processing so that the high temperature oxidation and diffusion times are not so short to be controllable. Suppose that the 100A thick oxide is to be grown at 900C in dry oxygen on a p-type Si.

(a) What is the oxidation time in seconds. (Answer: Near 300 seconds.)

(b) What is the capacitance of the oxide layer per unit area, C^o, in pF/cm^2?

P401.2 High voltage (100 to 400 Volts) and medium speed (1-20 MHz) MOS transistors and integrated circuits have been designed and volume produced in 1990 to drive non-impact printers (electrostatic or xerographic printers) and flat display panels (vacuum fluorescent displays used in automobiles and appliances, plasma display panels in computer monitors, electroluminescent displays), and in medical ultrasound imaging, diagnostic telecommunications, avionics, instrumentation, and industrial control equipment. The transistor dimension is larger to withstand the higher voltage and the gate oxide thickness is scaled in proportion and is about 500A.

(a) How long does it take to grow the oxide in dry oxygen at 1000C?

(b) How long does it take to grow the oxide in steam at 900C?

(c) What is the capacitance of the oxide layer per unit area in pF/cm$_2$? If the gate area is 5µm·500µm, what is the gate capacitance?

(d) What is the load resistance of the previous stage that drives the gate in order that the RC time constant of the gate meets the 20MHz speed requirement?

P402.1 The explanation given for the HFCV curve is that the minority carriers cannot be supplied to and extract from the oxide/silicon interface rapidly enough by diffusion or generation-recombination to respond to the signal voltage variation at 1MHz. Why is recombination also necessary and not just generation?

P410.1 Verify the negative sign in $C_s = -dQ_S/dV_S$ and $C_{it} = -dQ_{IT}/dV_S$ by physical argument without algebra via noting that increasing V_S more positively would increase N and decrease P.

P410.2 In this problem, use only (410.7), (410.7A) and $C_o = dQ_G/dV_o$, i.e., $C_g = dQ_G/dV_G = dQ_G/(dV_O+dV_S) = 1/[(dV_O/dQ_G)+(dV_S/dV_G)] = 1/[(1/C_o)+(dV_S/dV_G)]$. Assume the MOSC is truly one-dimensional and areally uniform.

(a) Show that the change of the surface potential between two applied d.c. gate voltage is given by following capacitance integral where C_g is the experimental capacitance.

$$V_S(V_{G2}) - V_S(V_{G1}) = \int_{V_{G1}}^{V_{G2}} \left[1 - \frac{C_g}{C_o} \right] dV_G$$

(b) Show that the above is valid at any frequency as long as the measured C_g contains no loss or resistive component.

(c) Show that the above is valid even when there is a large density of interface states which both distort the measured C_g-V_G curve and contribute a capacitance to the measured C_g via charging and discharging the interface traps at the signal frequency which is represented by C_{it}.

P410.3 The integral in problem P410.2 has been used by some researchers and engineers to compute the interface trap density of state in order to monitor an oxidation and a whole production process. Discuss the limitations of applying the above result to experimental C_g-V_G to compute the density of the interface traps given by C_{it} defined by (410.13), such as the uncertainty of the energy level of the interface traps.

P411.1 You can use the numerical values of the material parameters given in problem P411.2.

(a) Determine the area of the p-Si MOSC shown in Fig.411.1(b).

(b) Verify $N_{AA}=1.45\text{x}10^{16}\text{cm}^{-3}$ using the strong-inversion C_{HF}.

(c) Verify $N_{AA}=1.45\text{x}10^{16}\text{cm}^{-3}$ using the depletion C-V curve.

P411.2 The ratio of the inversion layer, x_I, to the very-strong inversion layer, x_{th}, was quoted as $\sqrt{2}/(\sqrt{2}-1)=1.414=0.414$ while discussing the space-charge distribution shown in Fig.410.1(b). Show how this ratio can be obtained from the depletion approximation.

P411.3 The MOSC area in the DRAM cell is 10 μm^2 and it is fabricated on an n-type Si substrate with a donor dopant concentration of $N_{DD}=1.0\text{x}10^{17}\text{cm}^{-3}$. The manufacturing technology is so refined and under control that there are no oxide and interface traps and that the work function difference between the gate conductor metal and the Si is zero by properly doping the polycrystalline silicon gate conductor. The chip is at room temperature. Assume $T=290\text{K}$, $kT/q=0.025\text{V}$, and $n_i=1.0\text{x}10^{10}\text{cm}^{-3}$.

(a) What is the oxide capacitance in pF and fF units? (1 pF$=10^{-12}$ Farad and 1 fF$=10^{-15}$ Farad). $\varepsilon_o=3.9\text{x}8.854\text{x}10^{-14}$ F/cm. (Answer: 50fF.)

(b) If this capacitance is used to store the one bit of information at a power supply voltage of 3.2V, how many electrons are stored on the capacitance? (Answer: even number, not by accident! Do you know why 3.2V is picked? Hint: SiO_2/Si barrier height.)

(c) Where is the Fermi level measured relative to the intrinsic Fermi level, E_F-E_I in eV? Give the result also in terms of the Fermi potential relative to the intrinsic Fermi potential, i.e. $V_P=$?

(d) What is the threshold voltage at the onset of strong surface inversion?

(e) What is the capacitance of the MOSC at the very strong inversion condition?

(f) What is the capacitance of the MOSC at the threshold or onset of inversion?

(g) What is the d.c. voltage at the intrinsic surface condition?

(h) What is the capacitance of the MOSC at intrinsic surface?

(i) What is the d.c. voltage at flat band?

(j) What is the capacitance of the MOSC at flat band?

P411.4 A parallel voltage shift of the experimental HFCV curve of the 100A-oxide MOSC in problem P401.1 was observed when compared with the ideal theory. The observed voltage shift was -0.1V.

(a) What is the density (Coulomb/cm^2) and sign of the oxide charge which causes this shift?

(b) What is the oxide trap density if each trap is singly charged?
[Formulae derivable from figures and equations in chapter 4: $Q_{OT}=-C_o\Delta V$ and $N_{OT}=|Q_{OT}/q|$.] (Answer: About 2×10^{11} oxide-trap/cm^2.)

P411.5 A distortion is observed in the experimental HFCV curve of the 100A-oxide p-MOSC in problem P401.1 after it is stressed in a very high electric field (about $V_G=8V$). The distortion is indicated by the non-equal voltage shifts observed at two capacitances, $C=0.9C_o$ and $C=0.8C_o$, and are $\Delta V_{G0.9}=+1.000V$ and $\Delta V_{G0.8}=+1.100V$. What is the interface state density generated during this stress and what is the range energy levels of these interface states?

P411.6 Assuming that the depletion condition exists in a MOSC on an n-type Si, obtain the expressions and sketch the electric field and the electric potential when V_G gives a large reverse bias to sustain the depletion condition. Refer to Figs.411.2(e) and (f) which give the curves for a MOSC on a p-type Si.

P411.7 Sketch to scale the electric field and electric potential at equilibrium like Fig.411.2(b), but assume $Q_{OT}=Q_{IT}=0$ in Fig.411.3(b). The curves are similar to Fig.411.2(e) and (f) but have significant difference which you must show in your figures.

P411.8 Can the depletion capacitance formula, (411.8), be obtained using strictly the circuit model of C_o in series with C_s to give the voltage V_S in terms of the total voltage V_G? Is your result identical to (411.8) and why?

P411.9 Derive the formulae for the low-frequency capacitance in the range of majority carrier (hole) accumulation in a p-Si MOSC that is stated in subsection 411(D) following the analytical procedure used in 411(C) for the low-frequency capacitance. Why is it valid to use the low-frequency model to compute the MOS capacitance in the majority carrier accumulation range?

P411.10 Derive the formula for C_p in place of (411.7) from the integral (410.9). Does your result reduces to (411.7) and under what condition?

P411.11 The flat-band capacitance and the Debye length concept described in subsection 411(E) for the p-Si MOSC are actually more general and apply to any region of a semiconductor. Demonstrate their validity in a p-type semiconductor region at equilibrium, labeled $x=0$, where the dopant impurity concentration varies with position, $N_{AA}=N_{AA}(0)+ax$ and the concentration gradient, a, is a constant. What is the restriction on 'a' in order that the validity is accurate? (Hint: The concentration gradient 'a' is much smaller than a characteristic-concentration divided by a characteristic-length.)

P411.12 It is evident from subsection 411(E) that the Debye length is a characteristic parameter of a semiconductor. Some semiconductor mathematicians and device physicists have used the intrinsic Debye length which is about 29μm in Si at 300K while the extrinsic Debye length is 410A at $N_{AA}=10^{16}$cm^{-3} computed after (411.17). It is evident that the intrinsic Debye length is not very meaningful. What is the intrinsic Debye length of Si, Ge, GaAs and GaP at 300K? Use the data of n_i from chapter 2. What is the Debye length (or extrinsic Debye length) of these semiconductors doped to 1.0×10^{16}cm^{-3} of donor impurities? Explain the usefulness of the extrinsic Debye length and the uselessness of the intrinsic Debye length.

P411.13 What is the Debye length in a heavily doped Si with donor impurity concentration of 3.0×10^{18}cm-3 at 300K and at 77K? Take into account the donor impurity deionization effect described in chapter 2.

P411.14 What is the high frequency capacitance of a p-Si MOSC in strong inversion and at 300K and 77K whose boron acceptor concentration is $3.0\times10^{18}\text{cm}^{-3}$? Take into account the acceptor impurity deionization effect described in chapter 2. Assume an oxide thickness of 100A and a $C_oA_{ox}=10$pF. How much error is introduced if impurity deionization is ignored?

P411.15 An accurate analytical high-frequency capacitance formulae, which has been extensively used in the Terman analysis of the experimental HFCV curves to study the MOSC properties such as Q_{OT} and Q_{IT}, was derived by this author in 1964 by partitioning the semiconductor surface space-charge layer into three layers: $x_d=x_1+x_2+x_3$ where x_1 is next to the interface, x_2 covers the potential range from $U_1 = 2U_F + \log_e(U_1-1)$ to $U_3 = 2U_F-2 + \log_e(U_3-1)$, and x_3 extends out into the bulk. The results are:
$x_1 = 2L_{DBi}\exp(-U_F/2)\{-\exp(U_F-U_S/2)+[e\surd(2U_F-1)]^{-1}\}$,
$x_2 = L_{DBi}\exp(-U_F/2)\{4+[\log_e(2U_F+1)]/(2U_F-3)\}/\surd2(U_F-1)$, and
$x_3 = 2L_{DBi}\exp(-U_F/2)\{2U_F-3+\log_e(2U_F-3)+[(2U_F-3)\exp(2U_F-2)]^{-1}\}^{1/2}$
where L_{DBi} is the intrinsic Debye length and U_F is Fermi potential normalized to the thermal voltage, $U_F=qV_F/kT$. Verify numerically the $2V_F+3kT/q$ approximation given in subsection 411(B) on the asymptotic high-frequency capacitance, $C_\infty \propto \surd[\varepsilon_s qN_{AA}/2(2V_F+3kT/q)]$.

P411.16 The 3kT/q additional potential variation used in the text and obtained in the preceding problem, P411.15, is illustrated in Fig.411.3(f) as all appearing at the interface while it is distributed in the entire space charge layer. How is it distributed or divided in the three layers described in the preceding problem?

P412.1 In the text, we described in words why the equilibrium high-frequency capacitance in strong inversion becomes a constant value much lower than C_o, such as those shown in Figs.411.1(a) and (b) for the n-Si and p-Si MOSCs. It was stated that $dQ_{MAJORITY}/dV_G \to 0$ as the inversion increases or $|V_G|$ increases in the inversion direction since most of the increasing $|dV_G|$ is compensated by attracting additional minority carriers to the semiconductor surface from the semiconductor bulk. Show how the C_S, $C_{MAJORITY}$ or C_{HF} in the strong inversion range can be computed using the algebra in section 412? Explain or note the meaning of the equilibrium condition in your discussion.

P412.2 Show that if we have a SOM instead of MOS structure, then the negative sign in $\phi_{MS}=-V_{BI}$ would drop out.

P412.3 Using the simple parallel capacitance formulae and not the complicated integration of the Poisson equation, show that Q_{OT}, in units of (C/cm^2), as defined by (412.13B) is actually the moment of the volume density (C/cm$_3$) of the oxide charge distributed in the oxide layer, such as that sketched in Fig.412.1(b). Where is the pivot point (line in this case) of the moment? Show that Q_{OT} gives the same amount of semiconductor energy band bending as a sheet of charge located at the oxide/semiconductor interface with the magnitude and sign of Q_{OT}. Show that $Q_{OT}=\int(x/x_o)\rho_{OT}(x)dx$ given by (412.13B) is less than the total integrated charge trapped in the oxide, $Q_{OTTOTAL}=\int\rho_{OT}(x)dx$.

P413.1 The contact transient example shown in Fig.413.1 illustrates electrons transferred from the metal gate to the p-type Si that is based on a specific set of assumptions on the material properties of the metal and Si. Construct the energy band for the case where electrons transfer in the opposite direction as Fig.413.1, i.e. from the p-type Si to the metal, during this contact transient. Indicate the condition on work function and electron affinity for this to happen. Try this also for an n-type Si.

P413.2 Using Fig.413.2 as a guide, construct the equilibrium energy band diagrams of an Al/SiO$_2$/n-Si MOSC at $V_G=0$, V_{FB}, V_{GI}, V_{TH}, $V_{INV}>>-V_{TH}=+|V_{TH}|$, and also at V_G=accumulation. Label the energy values in units of V_F. Use the parameters for Al, SiO$_2$ and Si given in Table 413.1 and other parameters in this chapter and chapter 2.

P413.3 Construct the depletion energy band diagram of the n-Si MOS capacitor of problem P413.2 in deep depletion.

P413.4 **Polycrystalline silicon** has been used as the gate material in Si MOS transistors. It is also used as a 'barrier' layer for making contact to thin Si layers to prevent a metal conductor from penetrating and short-circuiting the thin Si layer. But in MOS, poly-Si gate also gives one more parameter to control the threshold voltage (see next problem). Thus, construct the equilibrium energy band diagram of a polycrystalline silicon gate MOSC on n-type Si at $V_G=0$. Do this for four cases. (a) The Si-gate is doped to the same donor concentration at the n-type Si bulk. (b) The Si-gate is so heavily doped that its Fermi level is at E_C. (c) The Si-gate is doped to p-type with its acceptor concentration equal to the donor concentration in the n-Si bulk. (d) The Si-gate is doped to degenerate p-type with $E_F=E_V$.

P413.5 Compute the threshold voltage, V_{TH}, of the four cases given in problem P413.4 above assuming that $Q_{OT}=Q_{IT}=0$.

P413.6 Find the impurity concentrations necessary to dope the polysilicon gate so that the surface is flat-banded, $V_S=0$, intrinsic, $V_S=V_F$, and at the threshold, $V_S=2V_F$. The last is known as the zero threshold device but a more correct design is $V_S=V_F$.

P413.7 During the hypothetical contact transient of the three layers of the four MOS structures in problem P413.4, are electrons transferred from the gate conductor to the Si or vice versa? Explain the detailed sequence in the same way as Fig.413.1(a) to (f).

P413.8 What material parameter values are necessary so that at $V_G=0$, Al/SiO$_2$/Si MOSC has the condition of (a) flat-band, (b) intrinsic surface, (c) onset of strong inversion, and (d) in strong inversion with $V_S=2V_F+4kT/q$? If it is not possible to achieve some of these conditions using the Al gate, replace it with doped Si gate and determine the doping needed. These are important design conditions for MOST switches or inverters discussed in chapter 6. For examples, (b) would give zero threshold in a low-current circuit designed to operate in the subthreshold range, and (d) would give the depletion load transistor in the inverted circuit using an induced inversion channel instead of a doped channel which would require an additional ion implantation step.

P413.9 How many electrons are there in a Debye sphere of a metal? Express your solution in terms of the electron concentration. Then, find out at what electron concentration will the number of electron in a Debye sphere equal to one. Explain.

P413.10 Draw the equilibrium energy band diagrams similar to those given by Figs.413.2(a)-(c) if a sheet of oxide charge is present at x_{OT} from the gate-conductor/oxide interface. (a) $Q_{OT}>0$. (b) $Q_{OT}<0$.

P413.11 Under what combinations of ϕ_{MS} and a sheet of Q_{OT} located at x_{OT} from the gate-conductor/oxide interface will there be an oxide potential energy (a) maximum and (b) minimum.

P413.12 Follow Figs.413.3(a)-(f), draw the energy band diagrams under d.c. applied gate voltages for the six cases from $\phi_{MS}>=<0$ and $V_F><0$ (n-type and p-type).

P421.1 Use the parallel-series capacitance model and charge sharing among the capacitances. Derive the equation of the time dependence of the thickness of the semiconductor depletion layer, $x_d(t)$. Show that your result is identical to (421.2B) for bulk traps.

P421.2 Use the parallel-series capacitance model and charge sharing among the capacitances. Derive the capacitance transient due to a the presence of only a time varying oxide trap sheet located at a distance x_{OT} from the oxide/Si interface.

P421.3 A sheet of constant magnitude sodium ions, Na^+, is moving from the Al gate-conductor through the oxide towards Si. What is the capacitance transient waveform. Sketch a case in which the Na+ sheet is moving at a constant velocity. Sketch a second case in which the Na+ sheet is moving a constant electric field.

P421.4 Obtain as many material parameters as you can from the experimental phase I and II capacitance transient waveforms shown in Fig.420.1(c) to (e) for a gold-doped Si MOSC.

P421.5 A sheet of oxide traps is being charged and discharged via trap-band tunneling. What is the capacitance transient?

P422.1 A sheet of oxide traps is being charged and discharged via thermal trap-band transitions. Without doing any algebra and using symmetry alone, write down the capacitance and current transient formulae.

P422.2 A sheet of oxide traps is being charged and discharged via trap-band tunneling. Derive the equation of gate current waveform valid for all times assuming that this is the only time dependence.

P422.3 Calculate the Au bulk trap parameters from the phase I and II current transient shown in Fig.420.1(f) and (g). Assume a reasonable MOSC area, substrate doping, and oxide thickness based on examples given in the text or look up the values in the original article.

P423.4 An oxidized Si has an interface trap density of 10^{12}cm^{-2} located at the Si midgap with $e_{ns}{}^{t}=1000s^{-1}$ and $e_{ps}{}^{t}=1s^{-1}$. What is the peak (initial) current during the current transient observed on a p-Si MOSC ($N_{AA}=10^{16}$cm^{-3}) with an area of 10x10µm^2 and a oxide thickness of 100A? If this were a bulk trap uniformly distributed in Si, what would be its concentration in order to give the same peak current?

P430.1 Based on the physical considerations given in the text, draw out the complete small-signal equilibrium equivalent circuit model of a one-dimensional MOSC.

Chapter 5

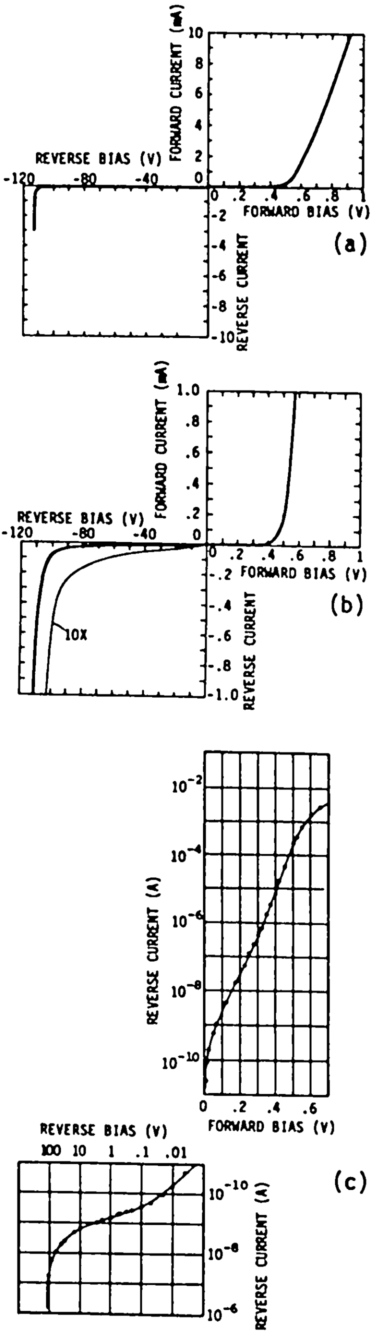
P/N AND OTHER JUNCTION DIODES

500 INTRODUCTION 382
510 FABRICATION OF A DIFFUSED P/N JUNCTION DIODE 385
    511 Diffusion and Fisk's Law, 385
    512 Physics and Data of the Diffusivity, 389
    513 Diffused Junction Depth Calculation, 395
520 EQUILIBRIUM ELECTRICAL PROPERTIES OF A P/N JUNCTION 397
    521 Equilibrium Energy Band Diagram, 399
        * Position of the Fermi Level, 399
        * Intrinsic Fermi Level and Electric Potential, 403
    522 Equilibrium Potential Barrier Height in a p/n Junction, 404
    523 Equilibrium Potential Variation in a p/n Junction, 408
    524 Application of the Gauss Theorem to a p/n Junction, 416
    525 Comparing the Depletion Approximation with the Exact Solution, 418
530 DC ELECTRICAL CHARACTERISTICS OF A P/N JUNCTION 420
    531 Energy Band Diagram of a Biased p/n Junction, 424
        * Reverse bias, 424
        * Forward bias, 427
    532 The Shockley Diode Equation, 430
    533 Physics of the Shockley Diode Equation, 432
    534 Numerical Example of a Shockley Diode, 436
    535 The Sah-Noyce-Shockley Diode Equation, 436
    536 Breakdown of Reverse DC Current in a p/n Junction, 441
        * Mathematical Formulation, 443
        * Basic Physics of the Parameters, 446
        * Simple Solution, 448
    537 Experimental-Theoretical Comparison, 451
540 SMALL-SIGNAL CHARACTERISTICS OF A P/N JUNCTION 455
    541 Small-Signal Charge-Control Circuit Elements, 456
    542 Small-Signal Numerical Examples of a Si p/n Junction, 458
550 SWITCHING TRANSIENTS IN A P/N JUNCTION 461
    551 Charge-Control Switching Analysis of a p/n Junction, 462
    552 Turn-On Transient in a p/n Junction, 464
    553 Turn-Off Transient in a p/n Junction, 467
    554 Capacitance and Current Trapping Transients in a p/n Junction, 470
560 METAL/SEMICONDUCTOR DIODE 474
    561 Equilibrium Energy Band Diagrams of the Schottky Barrier, 478
    562 DC Current-Voltage Characteristics of the M/S Diode - Bethe Theory, 483
    563 Experimental M/S Diodes, 489
    564 Effect of Semiconductor Voltage Drop - Mott Theory, 493
    565 Integrated Circuit Schottky Barrier Diode Layouts, 497
570 TUNNEL DIODES 499
580 LIMITING MECHANISMS OF DC TERMINAL CURRENT OF DIODES 500
    581 Current Limits in Metal/Semiconductor Diodes, 500
    582 Current Limits in p/n Junction Diodes, 502
    583 Contact Resistance, 506
590 SEMICONDUCTOR/SEMICONDUCTOR HETEROJUNCTION DIODES 510
    591 Energy Band Diagrams of s/s Heterojunctions, 510
        * Trapless s/s Interface, 510
        * Trappy s/s Interface, 512
    592 Electrical Characteristics of s/s Heterojunctions, 513
599 BIBLIOGRAPHY AND PROBLEMS 514

500 INTRODUCTION

The p/n junction is the principal element of a semiconductor p/n junction diode. Its unique electrical property is its highly nonlinear conductivity. It conducts current easily at low voltages when the d.c. voltage applied to the p-type terminal is positive relative to the n-type terminal and greater than about 0.5V. This is called the **forward bias polarity** and **forward current direction**. It conducts nearly no current even at moderately high d.c. voltages (5V or higher) when the applied voltage polarity is reversed (p-type terminal is negative relative to the n-type terminal). This is known as the **reverse bias polarity** and **reverse current direction**. This nonlinear electrical property is known as **rectification**. The nonlinear d.c. current-voltage characteristics or **I-V curves** of a historical (1957) silicon p/n junction diode is shown in Figs.500.1(a)-(c). The curves display the d.c. current **flowing into** the p-type terminal versus d.c. voltage **applied** to the p-type terminal **relative to** the n-type terminal. The structural view and the circuit symbol are given in the inset. Three expansions of the current scale and one expansion of the reverse voltage scale are shown. Figure (a) displays the linear I-V curve in the normal operation range in a circuit. Figure (b) shows the same linear I-V curve with two expanded current scales to demonstrate the I-V characteristics at lower currents. Figure (c) shows the I-V characteristics with further expanded current and voltage scales which was used in 1957 by this author to discover and delineate the fundamental controlling mechanisms of the magnitude of the current. In figure (c), the forward voltage is plotted on the same linear scale as figures (a) and (b) while the reverse voltage, forward current and reverse current are all plotted in the logarithmic scale. The key characteristic voltages are the forward and reverse voltages at which the current increases rapidly. They are shown in figure (a). It is about 0.5V in the forward direction, known as the **built-in voltage**, and -110V in the reverse direction, known as the **breakdown voltage**. Physical mechanisms responsible for the entire current-voltage characteristic are described in this chapter and used to derive the electrical characteristics and response under d.c., small-signal sinusoidal, and large-signal transient applied voltages or currents.

The nonlinear property is responsible for the many applications of the p/n junction diodes, for example, power rectifier to convert alternating current to direct current, signal detector in radio receivers, waveform clamp in digital circuits, photodetector, solar cell, light emitter, and many others. It has also been used as temperature and pressure sensors. The p/n junction is also the principal part of multi-junction/multi-terminal solid state amplifiers and switches, such as the emitter and collector junctions of bipolar junction transistors, the input terminal or gate of junction-gate field-effect-transistors, and the three junctions of the 4-layer two-terminal p/n/p/n diode switches and the 4-layer three- or four-terminal p/n/p/n silicon controlled power rectifiers. The p/n junction is also the essential element in modern silicon metal-oxide-semiconductor field-effect transistors (MOSFET), serving as the source and drain of electrons or holes and as the isolation to isolate the conduction path from the other transistors on the silicon integrated circuit chip.



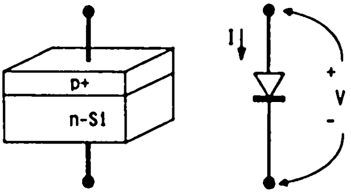


Fig.500.1 The current-voltage characteristic (I-V curve) of a historical (1957) silicon p/n junction diode shown in three expanded scales. (a) Normal operating current range, linear I versus linear V. (b) Expanded low current range, linear I versus linear V. (c) Greatly expanded current range used to discover the current limiting physical mechanisms, semi-log (logI-V) in the forward direction and log-log (logI-logV) in the reverse direction. (From Sah, Noyce, Shockley, 1957. See section 537.)

Semiconductor p/n junction is one of the general classes of junctions between two materials which have different chemical, physical and atomic makeups. Junction is the interface or boundary surface that separates the two materials. A solid-state junction may be a homogeneous junction or a heterogeneous junction. Homogeneous means that the materials on the two sides of the junction are identical chemically, physically and atomically. The material composition of the two sides of a p/n junction made on one species of semiconductor (such as Si or GaAs) is slightly different (in the one-part per million atom range) due to the presence of the minute amount of acceptor dopant impurity on the p-type side and donor dopant impurity on the n-type side. However, they are usually classified as homogeneous junctions or homojunctions. The boundary of a p/n junction is the interface through which the semiconductor changes from p-type to n-type. In contrast, heterogeneous junctions are made of two dissimilar materials which have significantly different physical and chemical or atomic makeups, such as a metal and a semiconductor in metal/semiconductor diodes; two semiconductors of significantly different atomic compositions (differing in the 10-100 atomic % range); and the insulator/semiconductor junction in the MOS capacitor studied in chapter 4.

A heterogeneous junction, known as a **heterojunction**, is the surface or areal contact, or the boundary between the two different materials. We will focus only on materials in solid form although heterojunctions between solid, liquid and gas are also present in electron devices and can affect the electrical characteristics of devices significantly. Non-solid and partially solid junctions are subjects of advanced courses.

Heterojunction made of a metal and a semiconductor is the oldest solid-state junction studied and used in applications. When a metal/semiconductor junction or contact has the rectification property, it is known as the **Schottky barrier**, the surface barrier, or the hot carrier diode. Schottky barrier diode is the oldest nonlinear solid state electronic device known to man (and woman). Its nonlinear property was used as the radio signal detector in the first radio transmission in the early 1930's, and as a power rectifier to convert alternating-electrical-current (a.c.) power source to direct current (d.c.). In the early 1940's during World War II it was used in radar receivers as the microwave signal detector. Later, in the 1960's, Schottky barrier diode was the key element that made the low-power high-speed TTL (bipolar transistor-transistor logic) IC (integrated circuit) possible. Because Schottky diode passes a much higher forward current than p/n junction diode, it is used to clamp the forward voltage across the collector p/n junction of the bipolar junction transistors during saturated switching operation. The forward voltage clamping prevents the collector junction from injecting a high density of minority carriers into the base layer, thus enabling the bipolar transistor to switch off quickly. Thus, the Schottky diode clamp speeds up the TTL logic gate circuits substantially.

Heterojunctions made of two compound semiconductors have been used to confine electrons, holes and photons in designated layers of a semiconductor device to produce unique device characteristics such as efficient and tunable semiconductor lasers by confining light in a thin layer of a few atoms thick (known as quantum well) using the different dielectric and optical properties of the different semiconductor layers. Multiple heterojunctions have also been used to confine the electrons and holes in layers of exceedingly high mobility in the hope of producing ultrafast transistors. Heterojunction has recently been used as the emitter junction of bipolar transistors to give very high current gain, $>10^3$, and cutoff frequencies, >100GHz or 10^{11}Hz.

If a solid-state junction conducts current equally well in both directions, it no longer rectifies and is known as an **ohmic contact**. Ohmic contact, although not electrically active and will not amplify, is a most important interface in solid-state devices. It enables low-resistance interconnection of diodes and transistors in monolithic integrated circuits on a silicon chip, and it provides the electrical contacts for off-chip external access of diodes and transistors.

In this chapter, we will study the electrical characteristics of p/n, metal/semiconductor junction diodes, ohmic and selected hetero junctions.

510 FABRICATION OF A DIFFUSED p/n JUNCTION DIODE

Modern Si p/n or n/p junctions are fabricated by diffusion, epitaxial growth, or ion implantation. They were made by a number of older technology no longer used today such as alloying (alloyed diode) and impurity doping during crystal growth (grown junction diode). In the following sections we shall describe the equipment and the physics of fabricating Si p/n junctions by solid-state diffusion of the impurity. Numerical example on junction depth calculation is also given.

511 Diffusion and Fisk's Law

Figures 510.1(a)-(c) illustrate the fabrication of a Si p/n junction by solid-state diffusion at high temperatures. Boron acceptor impurity is diffused into a phosphorus-doped n-type Si substrate to give the diffused p-type surface layer. The diffusion is carried out in a quartz tube in a furnace at a high temperature. The temperature is sufficiently high (1000°C) so that a junction depth of a few microns can be attained in a reasonably short time (tens of minutes) but not so high that the diffusion time is too short to be controllable. Diffusion produces a boron impurity concentration profile, i.e. concentration versus distance, illustrated in Fig.511.1(c). The silicon thickness is at least 300 microns or 12 mils to minimize breakage. In the current manufacturing practice of silicon integrated circuits, Si slices of 4-inch to 8-inch diameters are used whose thickness is 500-micron (20-mil) or thicker in order to prevent warping of the large-area wafer at high temperatures.

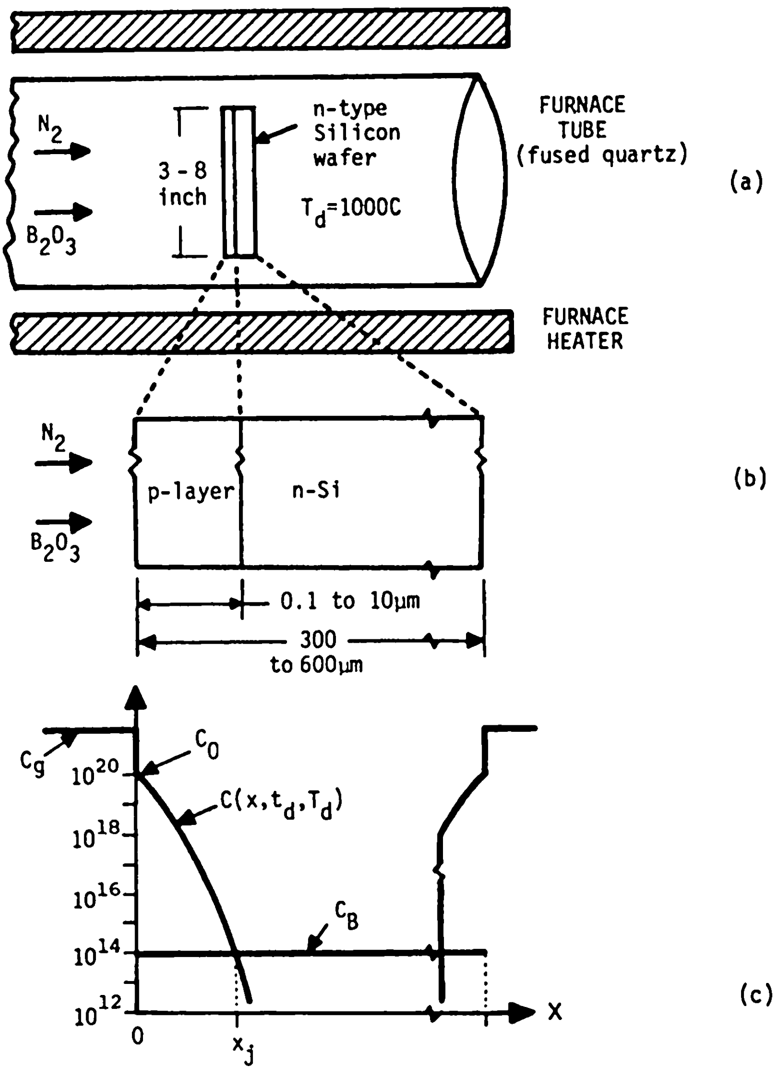

N2
B2O3
3 - 8 inch
n-type Silicon wafer
Td=1000C
FURNACE TUBE (fused quartz)
(a)
FURNACE HEATER
N2
B2O3
p-layer
n-Si
(b)
0.1 to 10μm
300 to 600μm
Cg
C0
C(x,td,Td)
CB
10^20
10^18
10^16
10^14
10^12
0
xj
X
(c)

Fig.511.1(a) Diffusion of boron impurity into an n-type silicon wafer in a furnace at 1000°C in flowing pure nitrogen containing a trace amount of B_2O_3 vapor. (b) Enlarged cross-sectional view in the thickness direction of the silicon wafer. (c) The diffused boron concentration profile in the Si surface layer, $C(x,t_d,T_d)$, and the constant (ideal) concentration, C_B, of the dopant donor impurity atoms (P, As, or Sb) originally present in the Si wafer.

For the example shown in Figs.511.1(a)-(c), the n-Si substrate has 10^{14} atom/cm^3 of phosphorus, arsenic, or antimony. Flowing through the diffusion tube shown in Fig.511.1(a) is pure and dry nitrogen with a trace of boron impurity in the form of boron oxide or B_2O_3 vapor. The partial pressure of the boron vapor is adjusted to give a surface concentration of 10^{20} boron/cm^3. Other forms of boron impurity source can also be used such as boron trichloride (gas), boron bromide (liquid), boron nitride (solid disks) and others. During the diffusion operation, the Si slice is placed in the fused-quartz tube inside a hot furnace maintained at about 1000°C shown in Fig.511.1(a). Boron in the nitrogen carrier gas stream will deposit or attach to the silicon surface. Since the boron concentration is high in the gas and low in the Si, boron atoms will move into Si by **diffusion**.

The position dependence of the concentration of boron, which is commonly known as the boron profile, impurity profile or **impurity concentration profile**, is governed by the **atomic diffusion equation** given by

$$\partial C/\partial t = D\partial^2 C/\partial x^2 \qquad (511.1)$$

where D is the **atomic diffusivity** of boron in Si at the elevated diffusion temperature and C is the concentration of boron. A typical boron concentration **profile**, i.e. C(x,t) vs x, is shown in Fig.511.1(c). It depends on both position and time, $C=C(x,t)$, where t is the diffusion time or the length of time at the diffusion temperature T during which the Si slice is exposed to the boron containing gas ambient. The above equation is a one-dimensional diffusion equation of the impurity atoms in a solid. It can be readily derived in the same way as that of the current and continuity equations of electrons or holes given by (350.1) to (350.4), but with two simplifications: (i) the drift current due to boron ion is negligible since the electric field is very small because the sample is essentially neutral and intrinsic owing to $N=P\approx n_i>>N_{AA}$ at the diffusion temperature; and (ii) generation, recombination and trapping of the boron impurity is negligible. These idealizations are modified in empirical theories used in engineering practice to design transistor fabrication process conditions.

In order to derive the diffusion equation from (350.2), the following replacements and modifications are made. The symbol C for the diffusant atomic concentration is used instead p for the hole concentration. Set q=1 so that j_P is the particle current instead of the hole electrical current since we are concerned here with the atomic or **particle current** of boron atoms rather than electrical current. Then (350.2) becomes the **continuity equation** of the atomic particles given by

$$\partial C/\partial t = -\nabla\cdot j \qquad (511.2)$$

and (350.4) becomes

$$j = -D\nabla C \qquad (511.3)$$

which is known as the Fisk's law. Here, j is the particle or boron atomic current density (boron/cm^2-sec passing through a plane located at x), D is the diffusivity of boron at the diffusion temperature, T_d, and assumed constant. Eliminating the particle current density j between these two equations gives

$$\begin{aligned} \partial C/\partial t &= \nabla\cdot j = \nabla\cdot(D\nabla C) \\ &= D\nabla^2 C \\ &= D[\partial^2C/\partial x^2 + \partial^2C/\partial y^2 + \partial^2C/\partial z^2]. \end{aligned} \tag{511.4}$$

The one-dimensional diffusion equation (511.1) is then obtained by dropping the y and z components in (511.4).

The solution of the one-dimensional diffusion equation, (511.1), is given by the complementary error function:

$$C(x,t) = C_0 \operatorname{erfc}[x/\sqrt{Dt}] \tag{511.5}$$

where

$$\operatorname{erfc}(Z) = 1 - \frac{2}{\sqrt{\pi}}\int_0^Z \exp(-x^2)dx = 1 - \operatorname{erf}(Z). \tag{511.5A}$$

Defined above, erfc(Z) is the complement of the error function, erf(Z). D is the diffusivity of boron at the diffusion temperature, $D=D(T_d)$. t is the diffusion time, t_d, but the subscript 'd' is usually dropped for notation simplicity. C_0 is known as the **surface concentration** or the concentration of boron on the surface of Si. It is sometimes denoted by C_S which tends to be confused with the substrate or bulk impurity concentration, C_B or C_{Sub}. C_0 is usually not too strongly dependent on the diffusion temperature or time but it does depend on the concentration of the boron in the gaseous ambient inside the furnace tube, C_g, shown in Fig.511.1(a). C_0 is usually smaller than C_g as indicated in Fig.511.1(c). The ratio, $C_0/C_g = k_g$, is known as the **distribution coefficient** of the boron impurity at the gas/Si interface. Notice that we have a similar parameter at the liquid and solid interface in the growth of Si crystals discussed in section 134. The distribution coefficient given here, k_g, has a similar meaning and effect as $k=C_{solid}/C_{liquid}$ in crystal growth. And, k_g is just as complex since there are just as many parameters and variables in gaseous diffusion, such as the temperature, gas mixture, gas streaming velocity, surface reaction rates and decomposition rates of the impurity molecules in the gas. It is a fundamental parameter in surface chemistry.

The complementary error function solution, (511.5), of the diffusion equation (511.1) was obtained using the following **initial** and **boundary conditions**.

$$C(x>0),t=0) = 0 \tag{511.5B}$$

$$C(x=0,t) = C_0 \tag{511.5C}$$

and

$$C(x=\infty,t) = 0. \tag{511.5D}$$

The initial boron concentration in the Si slice is zero as stated by the **initial condition**, (511.5B). The **first boundary condition** (511.5C) means that the boron concentration is kept constant on the Si surface during the entire diffusion. The **second boundary condition**, (511.5D), is approximate since the thickness of the Si slice is 300 to 600 microns and not infinite. It is a good approximation since the Si wafer is much thicker than the **diffusion length** of the boron impurity, $\sqrt{Dt}$, which appears in (511.5). Thus, the boron that is diffused into the back surface does not reach the front surface as indicated in Fig.511.1(c). For example, at 1000°C, the boron diffusivity in Si is 0.02 micron^2/hour and the boron penetration into the Si surface in one hour is ~ $\sqrt{Dt}=\sqrt{0.02}=0.14\mu m$ which is very small compared with the Si wafer thickness (~$500\mu m$) as indicated in Fig.511.1(c). Thus, when we solve (511.1) for the boron profile at the front surface, we can neglect the boron diffusion from the back surface through the Si wafer to its front surface. This lets us move the **theoretical back surface** to $x=\infty$.

Note that we are not using physicists' and chemists' unit of cm^2/s for diffusivity in solids which was used for electrons and holes at room temperatures. Instead, we use a much smaller engineer's unit, μm^2/hr, since boron (and all other impurities) does not diffuse very rapidly and its diffusivity is extremely small even at 1000°C. This unit was proposed by Shockley in 1959 when he started a transistor manufacturing company. It is precisely this rather small diffusivity of impurity atoms that allows us to accurately control the thickness of very thin p- and n-type layers in the manufacturing of high-speed and high-frequency p/n junction transistors.

512 Physics and Data of the Diffusivity

C(x,t) is also a function of the diffusion temperature T_d. The temperature dependence comes from the temperature dependence of the diffusivity of the impurity atom, $D(T_d)$, which is given by the Arrhenius equation

$$D(T_d) = D_0\exp(-E_A/kT_d). \tag{512.1}$$

E_A is known as the **thermal activation energy** and D_0 the **pre-exponential factor**. They are the fundamental parameters characterizing a diffusing atom in a particular solid or material. Their magnitude is determined by the properties of the two pathways in the host lattice along which the impurity atoms migrate: the **substitutional pathway** and the **interstitial channel** or tunnel. We shall avoid using the term 'tunnel' since it is reserved for quantum mechanical tunneling through a potential barrier.

In the interstitial pathway, the potential barrier is so low that the atoms are likely to jump over the barrier instead of tunnel through the thick barrier from one interstitial site to an adjacent interstitial site. Furthermore, there are probably few if any sites on the wall of the interstitial channel that will trap the interstitial atom. Thus, in the interstitial diffusion, the impurity atom moves through essentially 'empty' spaces or channels in-between the covalent bonds. E_A is then small in interstitial diffusion due to low intervening barriers and few binding sites.

However, along the substitutional pathway, the bound impurity atom must jump to the neighboring host (Si) vacancy over the high potential barrier from the intervening Si host ions. Thus, substitutional diffusion involves two steps, breaking the impurity-Si bond and jumping over the intervening barrier to the neighboring vacancy. E_A is then expected to be large in this vacancy mechanism of substitutional impurity diffusion.

The substitutional pathway is readily illustrated by a two-dimensional lattice shown in Fig.512.1(a). However, the channels of the interstitial diffusion do not show well in the two-dimensional lattice picture given in Fig.512.1(b). A better view can be demonstrated in the class lecture by the instructor using the three-dimensional ball-and-stick model of a crystal. For the diamond lattice of the Si crystal shown in Fig.132.2(a), the main interstitial channel is a squeezed hexagonal cylinder in the <110> direction with four 2.35A adjacent sides and two 1.43A facing sides, and a large 3.8A opening. The second channel is in the <100> direction with a square cross-section of 1.5A sides which is too small for fast interstitial diffusion of the large impurity atoms.

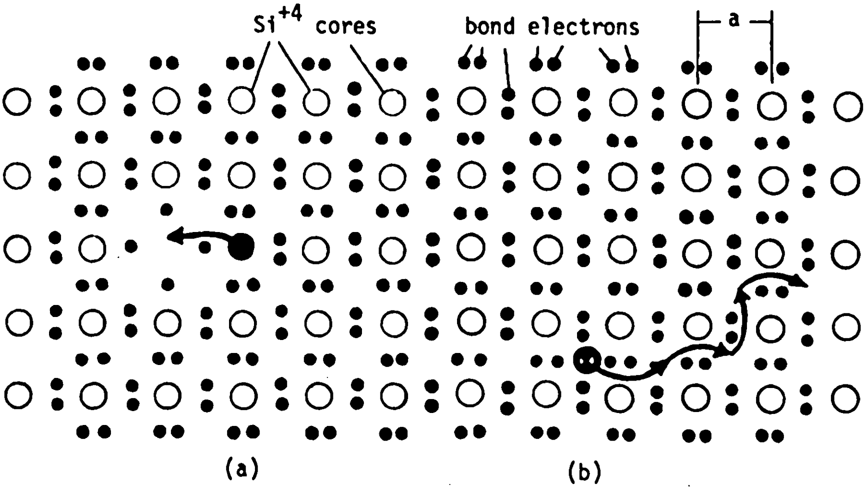


Fig.512.1 Atomic models for impurity diffusion. (a) Substitutional diffusion via the vacancy mechanism. (b) Interstitial diffusion.

The physical model just described is consistent with experimental data. Boron, phosphorus and the rest of the group-III and group-V impurities diffuse in Si mainly via the vacancy mechanism of substitutional diffusion with a D_0 of the order of 1-10 cm^2/s and $E_A=3.5$-4 eV. Au, Zn, Co, Li and other impurities diffuses in Si mainly via the interstitial channel with much smaller D_0 (10^{-2} to 10^{-3} cm^2/s) and E_A (0.6-0.7eV). However, the interstitial diffusivity is substantially higher than the substitutional diffusivity in spite of smaller D_0 because E_A dominates in $D=D_0\exp(-E_A/kT)$, and E_A is much smaller in interstitial diffusion.

Trapping of the diffusing impurities by Si vacancies limits the particle current. In computer aided design of integrated circuits, it is often not separately treated like trapping of electrons and holes in the Shockley equation. Instead, it is lumped into the diffusivity by making D_0 and E_A empirical fit parameters.

Physics of the Diffusivity

A simple physical model can be readily developed for the vacancy mechanism of substitutional diffusion that accounts for the magnitude of both D_0 and E_A. There are three fundamental steps during one migration jump of a substitutional impurity: (i) the creation of a vacancy, mainly on the surface or interface of Si where there are many Si with dangling bonds, (ii) the migration of the vacancy towards an impurity to sit next to the impurity (creating a vacancy-impurity pair), and (iii) the jump of the impurity into the vacancy leaving a vacancy behind. Thus, D_0 and E_A both consist of three parts. For example, E_A consist of (i) the formation energy of the vacancy at a surface or interface which is roughly equal to the energy of breaking the four covalent bonds around one Si surface atom at the diffusion temperature which is $4E_G=4\text{x}0.52\text{eV}=2.08\text{eV}$. The energy gap of Si at T=1300K is computed from

$$E_G = 1.21 - 4.15\text{x}10^{-4}T - 1.37\text{x}10^{-9}\sqrt{(n_i/T)} \qquad (512.2)$$

which gives

$$E_G = 1.21 - 0.54 - 0.15 = 0.52\text{eV}, \qquad (512.3)$$

(ii) the energy to move the vacancy into a site next to the substitutional impurity which is relatively small, and (iii) the energy for the impurity to jump into the adjacent vacancy which varies with the impurity core size and charge but is about the energy to break three (why?) covalent electron-pair bonds which is $3E_G = 3\text{x}0.52 = 1.56\text{eV}$. Thus, the total thermal activation energy for substitutional impurity diffusion via the vacancy mechanism is about $7E_G=3.64\text{eV}$. The experimental values in Si are empirical fit parameters but they are fairly close to the simple model estimate. The data are: B(3.6eV), Al(3.4eV), Ga(3.5eV), In(3.6eV), Tl(3.7eV) P(3.66eV), As (4.1eV), Sb^0(3.65eV), Sb^-(4.08eV).

The pre-exponential factor D_0 is mainly determined by the jump rate of the impurity into a nearby vacancy. It can also be estimated easily based on the simple picture shown in Fig.512.1(a). It is proportional to the jump probability rate of an impurity atom moving into one of the neighboring Si vacancies and the random fluctuation of the jump distance. It can be estimated from $D_0 = <n_{site}(\Delta x)^2\nu>$. Here, n_{site} is the number of substitutional sites surrounding the substitutional impurity which can be occupied by a vacancy. For the diamond lattice, there are 4 nearest neighbors, 12 second nearest neighbors, 36 second nearest neighbors, 108 third nearest neighbors, etc. Δx is the jump distance. For Si, Δx for the neighbor sites are: nearest neighbors (2.53A), second neighbors (4.2A), third neighbors (6.9A). Since the probability of a jump is proportional to the number of final site but is smaller for longer jump due to thicker barrier and larger distance, we make a rough estimate by taking $a=5.43A$ for the cubic unit cell of Si and $n_{site}=8$ for the eight lattice sites in this cubic unit cell. ν is the frequency of the lattice vibration which causes the jump and supplies the required energy to move the impurity. At the diffusion temperature of about 1000C or 1300K (bright red hot meaning there are many infrared optical phonons), the average kinetic energy of the lattice vibrations or phonons is about $KE = 3kT/2 = (3/2)25.85meVx(1300/300) = 168meV$ which is 2.6 times greater than the maximum optical phonon energy of the Si host lattice, which is 64meV (or $\nu_{max}=1.6x10^{13}s^{-1}$) for the optical phonon at $q=0$ shown in Fig.313.3(a). Thus, there are many high energy phonons to cause the jumps and the jump frequency can be taken as ν_{max}. Using these estimates, then for substitutional impurity diffusion,

$$D_0 = <n_{site}(\Delta x)^2\nu> \simeq 8a^2\nu_{max} \tag{512.4}$$

$$\simeq 8(5.43x10^{-8}cm)^2 1.6x10^{13}s^{-1} = 0.377\ cm^2/s. \tag{512.5}$$

Experimental values of D_0 in Si are: B(2.6cm²/s), Al(1.4), Ga(3.6), In(0.79), Tl(1.37), P(3.85), As(13), Sb^0(0.22) and Sb^{-1}(13). They have very large uncertainties since fitting the temperature-dependence data to the equation $D_0\exp(-E/kT)$ inevitably gives very large error in D_0 unless the data points are extremely accurate and cover many (five to ten) decades.

Diffusivity along the interstitial pathway or channels is more difficult to estimate based on simple physical considerations since the potential barrier separating the two adjacent interstitial sites in the channel is not known and the binding energy at the trapping sites on the interior surface of the channel is also not known. However, it should be much less than the energy required to move the group-III and group-V impurities substitutionally which was taken as 7 times the electron-pair bond energy (0.52eV at the diffusion temperature). In fact, it should take a smaller energy than that required to move an impurity core to its nearest neighbor vacancy (or flip an impurity-vacancy pair by a 180° rotation) which was taken as 3-times E_G or 1.56 eV at 1000C. The experimental data of interstitial diffusion have large scatter due to many interstitial pathways which are highly dependent on the crystal perfection. The data show an activation energy of about 0.6-0.7 eV for many interstitial impurities and a pre-exponential factor of 10^{-2} to

10^{-3} cm^2/s. If we assumed that there is only one adjacent site, then D_0 can be readily estimated since the interstitial site separation, a_i, is about 4.1A in Si giving an interstitial pre-exponential factor of

$$D_0 = \langle n_{site}(\Delta x)^2 \nu \rangle \approx a_i^2 \nu_{max} \tag{512.6}$$

$$= (4.11\times10^{-8})^2\times1.6\times10^{13} = 0.027 \text{ cm}^2/\text{s} \tag{512.7}$$

This estimate lies in the range of the empirical experimental data of D_0(cm^2/s) and E_A(eV) in Si which are: Ag(2×10^{-3},1.6); Au(2.4×10^{-4},0.39); Co(2×10^{-2},0.69); Cr(1×10^{-2},1.0); Cu(4.7×10^{-3},0.43); Fe(1.0×10^{-3},0.68); Mn(7×10^{-4},0.63); Ni(2×10^{-3},0.47); and Li(2.5×10^{-3},0.655).

Diffusivity Data in Si

The experimental diffusivity data of groups III and V substitutional impurities and oxygen in Si are graphed in Fig.512.2 as $\sqrt{D}$(μm/$\sqrt{}$hour) vs T°C. The Arrhenius plot of $\sqrt{D}$ vs 1000/T°K of these and metallic interstitial impurities are given in Fig. 512.3. The Arrhenius plot shows the nearly straight line behavior of a thermally activated process whose slope is the thermal activation energy E_A.

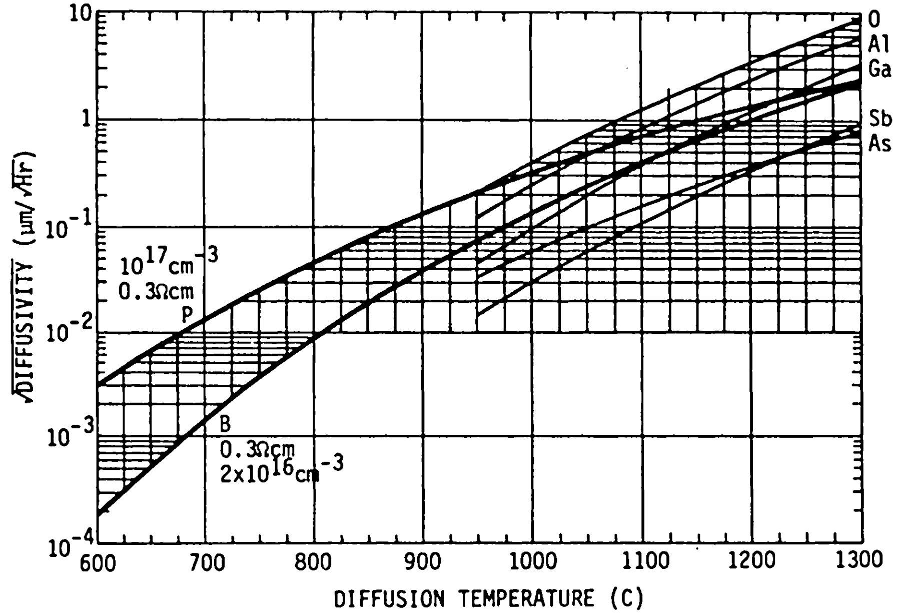


Fig.512.2 Temperature dependence of the diffusivity of impurities in Si. $\sqrt{D}$(μm/$\sqrt{}$hour) vs T(C°).

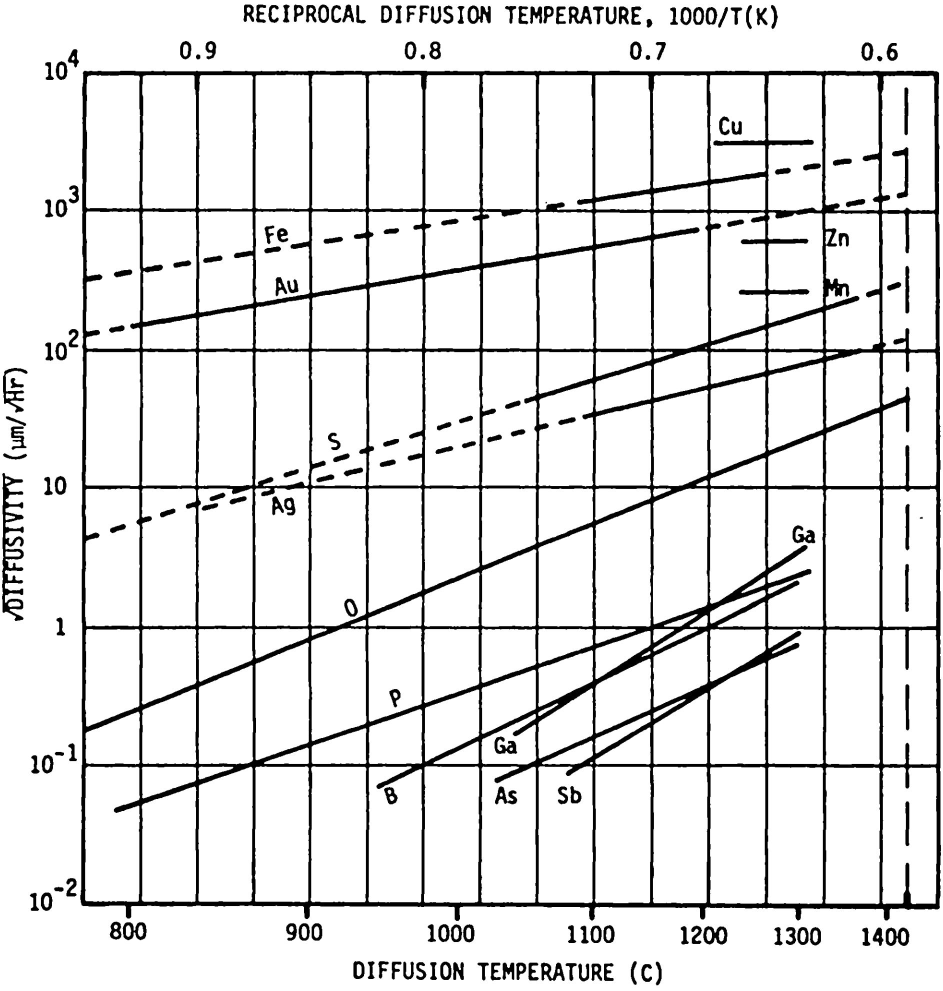


Fig.512.3 Arrhenius plot of impurity diffusivity in Si. √D(μm/√hour) vs 1000/T(K).

513 Diffused Junction Depth Calculation

Generally, once a C(x) profile is formed at the diffusion temperature, it is fixed and will not change at low temperatures such as 300°K or 25°C where the transistor or diode is operated at. Thus, in device analysis, we seldom write $C(x,t_d,T_d)$. Instead, we use the simplified notation C(x) or $N_{AA}(x)$ and $N_{DD}(x)$. The immobility or low mobility of the impurity atoms in a semiconductor at room temperatures is the fundamental reason that the characteristics of diodes and transistors do not change with time and electronic-electrical equipment can operate for ten or more years without failure. It is this very immobility of the impurity atoms at room temperatures that enables us to treat the semiconductor containing an atomically inhomogeneous impurity layer as if it is at **atomic** (or **chemical**) and **ionic** (or **electrochemical**) **equilibrium**.

The **p/n junction boundary** is defined as the position where the donor dopant impurity concentration, N_{DD}, is exactly equal to the acceptor dopant impurity concentration, N_{AA}, $N_{DD}=N_{AA}$. This is denoted by the symbol, x_j. Different symbols are sometimes used by other authors. The position of x_j is illustrated in Fig.511.1(c) which has $N_{AA}=N_{AA}(x)=C(x)$ and $N_{DD}=C_B=$constant.

This p/n junction boundary is perhaps more correctly called the **metallurgical junction**. It is a rather ancient name which was invented for **alloyed junction** diodes in which a piece of metal containing a p-type dopant impurity is alloyed into an n-type germanium at a high temperature. In this textbook, we will call this the **p/n junction boundary**. It is defined as the position where $N_{AA}(x)=N_{DD}(x)$ and labeled by $x=x_j$. We also use 'metallurgical junction' occasionally as a historical reminder.

The position, x_j, also defines an **operational** or **design thickness** of the p-type boron diffused layer as illustrated in Fig.511.1(c). It is frequently called the **junction depth**. The word depth refers to how deep the boron has diffused into and penetrated below the Si surface. Mathematically, we can compute the junction depth if we are given C_0, C_B or N_{DD}, D and t. The solution is given by

$$N_{AA}(x=x_j) = C_0\mathrm{erfc}(x_j/\sqrt{2Dt}) = N_{DD} = C_B \tag{513.1}$$

which can be solved to give the junction depth

$$x_j = 2\sqrt{Dt}\ \mathrm{erfc}^{-1}(C_B/C_0). \tag{513.2}$$

In order to solve (513.2) for a given Dt, C_0 and C_B, we need to use a table of complementary error function or error function which may not be handy. Thus, alternatively, we will use a simplified solution which enables us to compute numbers with a slide rule or a hand-held calculator. But it corresponds to a more complicated boundary and initial condition encountered more often in practice. This simplified solution is the Gaussian impurity distribution given by

$$C(x) = C_0 \exp(-x^2/4Dt) \tag{513.3}$$

where $C_0 = Q/\surd(\pi Dt)$ is no longer constant but a function of the diffusion temperature due to $D = D(T_d)$, and diffusion time, $t = t_d$. Solution (513.3) comes from the condition that the total boron atoms per unit area is a constant at all times and given by Q (atom/cm^2). It corresponds to the following practical sequence. First, a thin layer of Q boron/cm^2 is deposited into a thin Si surface layer by exposing to 10keV boron ions, known as **ion implantation**. Or it may be **prediffused** into the Si surface at a low temperature. Then the deposited or prediffused boron layer is **driven** deeper into the Si by diffusion at a higher temperature, T_d, for a time t. During this high-temperature diffusion, the surface of the boron layer is sealed so that boron cannot escape into the ambient and can only diffuse into Si. This theoretical assumption is nearly met in practice. After this **drive-in** diffusion, the boron concentration profile is given by (513.3). The junction depth from $C(x_j) = C_B$ is then

$$x_j = 2\sqrt{Dt \cdot \log_e(C_0/C_B)}. \tag{513.4}$$

Numerical result can now be obtained using slide rule or handheld calculator without having to find an error function table. Let t = 1.0hour, D = 0.02μm^2/hour, $Q = 2.5\text{x}10^{15}$B/cm^2 (determined apriori by the required sheet conductance of the diffused layer), and $C_B = 10^{14}$cm^{-3}, then

$$C_o = Q/\sqrt{\pi Dt} = 2.5\text{x}10^{15}/\sqrt{\pi \text{x} 0.02 \text{x} 1} = 10^{20}\text{cm}^{-3} \tag{513.5A}$$

$$x_j = 2\sqrt{0.02 \cdot \log_e(10^{20}/10^{14})} = 1.05 \text{ micron} \tag{513.5B}$$

If this junction is too deep or the diffused layer too thick for high speed transistors, then, we reduce the diffusion time to decrease the junction depth or the layer thickness. Equation (513.4) shows that the junction depth is roughly proportional to the square root of the diffusion time, t. It is not exactly proportional because the surface concentration will decrease with diffusion time since $C_0 = Q/\surd(\pi Dt)$ and Q = constant. This lowering C_0 will reduce the junction depth. However, the junction depth is not a strong function of C_0 compared with its square root dependence on the diffusion length $\surd Dt$ since (513.4) shows that C_0 is in the argument of the logarithmic term. However, the diffusion time cannot be reduced too much because D(T) cannot be predicted accurately since the heating and cooling times of the Si wafers and wafer carrying boat cannot be controlled reproducibly. Thus, if still thinner or shallower junctions are required, it is better to reduce the diffusion temperature rather than diffusion time since the diffusivity decreases rapidly with lowering diffusion temperature owing to the large thermal activation energy ($E_A \approx 4$eV) in $D = D_0 \exp(-E_A/kT)$. For example, Fig.512.1 shows that $\surd D_{boron}$ drops by a factor of about 10 when the diffusion temperature is lowered by about 100C from 1000C to 900C, giving $x_j = 0.1\mu$m or 10x shallower junction and thinner layer.

520 EQUILIBRIUM ELECTRICAL PROPERTIES OF A P/N JUNCTION

In order to compute the d.c. steady-state current-voltage characteristics and small-signal and large-signal dynamic responses of a p/n junction we need first to understand its characteristics at **thermal** as well as **electrical** or more restrictively, **electronic equilibrium**. We have assumed that when the p/n junction is operated in the room temperature range, it is at **atomic** or **chemical equilibrium**, that is, its atomic structure does not change or its atoms do not migrate during the operation period. The atomic cores only **vibrate** about their equilibrium position to conduct, store, and supply the electron kinetic energy. By **thermal equilibrium**, we mean that the p/n junction is at a constant temperature which is independent of position and time. To maintain this space-time constant temperature, the device is shielded from applied electromagnetic and ambient molecular disturbances, such as a spatially localized steady-state or pulsed electromagnetic radiation in the infrared range or a gas stream, which can heat up portions of the device structure to produce a temperature gradient. If we had a temperature gradient, a voltage will appear across the two terminals of the p/n junction. This voltage is known as the thermo-electromotive force or thermal emf and it is the voltage source of a thermocouple. By **electrical equilibrium** we mean that the electrical currents carried by electrons, holes and any atomic ions which could migrate are each equal to zero. Since we have already assumed atomic or chemical equilibrium at the normal operating temperatures, then migration of impurity ions is already assumed to be negligible. Thus, the additional conditions from assuming electrical equilibrium are those of **electronic equilibrium** which are the conditions of **zero electron current** and **zero hole current**. To maintain zero electron and hole currents, the p/n junction cannot be exposed or coupled to an optical or radio-frequency electromagnetic field and its terminals are not connected to an applied voltage source so that both its terminal current and its terminal voltage are zero.

If we consider further and look deeper into the basic perturbing forces that can cause nonequilibrium, it becomes immediately evident that the cause of thermal and electronic nonequilibrium is of the same origin: electromagnetic disturbances, including even the thermal nonequilibrium caused by a jet stream of ambient gas molecules. The difference is the wavelength, frequency or energy of the photons and the corresponding electromagnetic waves or fields. Thermal equilibrium deals with photons in the infrared (or heat) range which is also the fundamental cause of hot (or cold) ambient and the jet stream of ambient gas molecules that bombards the semiconductor surface. Electronic equilibrium concerns not only the lower-frequency electromagnetic waves or fields, and voltages or currents, but also the higher-frequency range in the optical and visible range where the photon energy is comparable to energy gap of the solid.

Strictly speaking, the p/n junction is not at thermodynamic equilibrium (containing all the equilibria or equilibrium of all species of particles) because of the inhomogeneous distribution of the donor and acceptor impurity atoms which puts the p/n junction at an atomic or chemical nonequilibrium condition. But, at room and normal operating temperatures, the diffusivities of the dopant impurity atoms are so small that they are essentially fixed in the Si lattice. One can readily illustrate this immobility by an empirical extrapolation of the high temperature diffusivities of the boron or phosphorus impurities in Si to 300K. Ignoring the increasing activation energy with lowering temperature, a linear extrapolation of the Arrhenius plot of phosphorus in Fig.512.2 to 300K gives $\sqrt{D} \approx 10^{-15}$ μm/$\sqrt{}$hour $\approx$ 10^{-9} A/$\sqrt{}$year or it will take 10^{18} years for a phosphorus donor to move about 1 A in crystalline Si at room temperature. A similar estimate for Au in Si gives 20A/$\sqrt{}$year (too high due to increasing E_A at lower T). Thus, **atomic, chemical or structural equilibrium**, i.e. **quasi steady-state**, is an excellent assumption within the period of ten or twenty years of the operating life of a p/n junction diode.

At least four characteristic times relevant to the various equilibria or (dynamic) quasi steady-state conditions can be defined for an atomically nonuniform solid, such as a semiconductor p/n junction. (i) The **microscopic time** between successive electron and hole scattering events, of the order of 10^{-13} second, to reach the thermal equilibrium distribution of the electron and hole kinetic energy. (ii) The **macroscopic time** ranging from 10^{-13} to 1 second due to the diffusion-drift-generation-recombination-trapping-tunneling delays of an ensemble of electrons and holes (or signal carried by a group of electrons or holes) while making the transit through a device layer which covers the entire time or speed range in circuit application. (iii) The **reliable operation time** ranging from a few seconds to 10^9 seconds, which is about 20 years and is the targeted operating life of transistor and integrated circuit. (iv) The **structure deterioration** or **homogenization time** covering the range in (iii) and extending to longer times when atomic density nonuniformity may begin to even out or homogenize by atomic diffusion and drift causing the device structure to disappear. The last, (iv), the structure deterioration time, has not been a concern for Si transistors for two reasons: (i) they are relatively homogeneous on the atomic scale and (ii) the dopant impurity density inhomogeneities are incorporated at high temperatures while the devices operate at the much lower (room) temperatures. However, it is a very serious fundamental limitation in highly atomic heterogeneous device structures, such as the compound semiconductor and Ge_xSi_{1-x} heterojunction transistors whose hetero-interfaces are fabricated at low temperatures.

The equilibrium internal electrical characteristics of a p/n junction will be derived first. They can then serve as the base to determine the effects of an applied d.c. or time-dependent voltage and current on the internal electrical characteristics and circuit response of the p/n junction. Thus, the remaining sub-sections, 521-525, will be completely devoted to a study of the electric potential, electric field and electric charge distributions in a one-dimensional p/n junction at thermal and electrical equilibrium. Considerable efforts will be devoted to describing the physical bases of the mathematical methods and the derivation procedures.

521 Equilibrium Energy Band Diagram

The E-x (energy-distance) energy band diagram of a p/n junction is an extremely useful visual aid in discussing and understanding the properties of a p/n junction and more complicated semiconductor devices. An energy band diagram can be used to visually illustrate the diffusion, drift, generation, recombination, and trapping of electrons and holes. It can also help to delineate the mechanisms which limit the current flowing in a semiconductor device. Furthermore, it is particularly useful in formulating the mathematical theory of the electrical properties of a device by helping to make the correct approximations based on physical intuitions and insights. Making approximations and making the correct approximations are the essential expertise for deriving analytical solutions of the p/n junction current-voltage characteristics since we cannot solve the six Shockley equations analytically or even numerically within a finite time at high accuracy on a supercomputer if it is a three-dimensional device structure. Furthermore, numerical solutions do not provide a physics base to enable us to understand and delineate the important parameters that determine the electrical properties of a p/n junction and more complicated devices. We have already demonstrated the usefulness of the E-x energy band diagram in the derivation of the MOS capacitance characteristics, such as the built-in potential in (412.2) and section 413, and the capacitance and current transients in sections 42n. Nevertheless, we will introduce the energy band diagram again in this section on p/n junction for those who did not follow through these advanced MOSC sections.

To illustrate the use of the energy band diagram, we shall employ a simple one-dimensional p/n junction whose dopant impurity concentration is spatially constant on both side of the junction. It is illustrated in Fig.521.1(b). The p-type side has a constant boron concentration of N_{AA} from $x=-T_p$ to 0 where T_p denotes the thickness of the p-type side. Similarly, the n-type side has a spatially constant phosphorus concentration of N_{DD} from $x=0$ to $+T_n$.

The step-by-step procedure of drawing the equilibrium E-x energy band diagram is now described because the physical reasoning behind each step helps to formulate the mathematical approximation later. Equilibrium means: the voltage applied to the two terminals of the p/n junction diode is zero; the diode is shielded from light; and the diode is at a uniform and constant temperature by being in perfect thermal contact with and immersed in a perfect (infinite size and heat conductivity) heat bath.

Position of the Fermi Level

Let us first compute the Fermi level position at the two end contacts of the diode since the Fermi level is spatially constant at equilibrium, as proved in section 330 which let us draw a horizontal line at the middle of the graph paper. Figure 521.1(c) shows the first step: draw a horizontal line for the Fermi level E_F.

Figure 521.1(d) shows the second step. Here we use the physical intuition that at positions far away from the p/n junction boundary ($x=0$), the energy band diagram should approach that of a neutral material, i.e., $\rho=0$ at $x=\pm\infty$. Furthermore, since we have assumed that N_{AA} and N_{DD} are both constant (but not necessarily equal), and that there is no current, the electric field E must also be zero far away from the p/n junction boundary, $x=0$. $E=0$ indicates that the energy levels in the energy band diagram should be horizontal or flat. This is known as **flat band** just like that of the MOSC. Thus, we draw horizontal E_C, E_V and E_I lines at both terminals of the p/n junction as shown in Fig.521.1(d). E_I is the intrinsic Fermi level and is used as the reference for the potential energy. Although we used the vacuum level as the reference when we discussed the energy bands in semiconductors in Chapter 1 (section 172), we shall use $E_I(x)$ for the electric potential in device analyses which is defined by $V_I(x) = E_I(x)/(-q)$. A convenient point x_{ref} is selected to set the reference potential to zero, $V_I(x_{ref})=0$. The choice of x_{ref} is made to put the solution into the simplest algebraic form so physics is not obscured by mathematics. Hence, x_{ref} can be different for each device property being analyzed in one device. V_I has been known as the **electrostatic potential** which was coined by Shockley in 1949 while analyzing the d.c. biased p/n junction which had no time dependence and hence the word '**electrostatic**'. Generally, the potential will vary with time if a time-dependent voltage is applied to the two terminals of a p/n junction diode. Thus, the more appropriate term is **electric potential** and symbol, $v_I(x,t)$, following the IEEE symbol convention.

The position of the Fermi level, E_F, relative to E_C, E_V (conduction and valence band edges) and E_I (intrinsic Fermi level) can be computed using the charge neutrality condition and the relationship between the Fermi energy and the electron and hole concentrations at equilibrium. These relationships are given by (242.11) to (242.14). Two computation methods are illustrated next.

Let us first compute E_F-E_V labeled (2) and E_C-E_F labeled (3) in Fig.521.1(d). We assume that both the n-type and p-type sides are extrinsic, that is, $N_{DD}>>n_i$ and $N_{AA}>>n_i$. Then, we can use the approximations (242.11A) and (242.13A) to compute the electron and hole concentrations.

$$N(x=+T_n) \equiv N_N \simeq N_{DD}(+T_n) - N_{AA}(+T_n) = N_{DD} \tag{242.11A}$$

and

$$P(x=-T_p) \equiv P_P \simeq N_{AA}(-T_p) - N_{DD}(-T_p) = N_{AA}. \tag{242.13A}$$

T_n and T_p are the thickness of the n-type and p-type layers. Using the Boltzmann approximation of the hole concentration given by (233.10) at the p-type terminal, $x=-T_p$, [labeled (2) in Fig.521.1(d)], then

$$P = P_P = N_V\exp[-(E_F-E_V)/kT] \simeq N_{AA}$$

and

$$E_F - E_V \text{ (at } x=-T_p) = kT\log_e(N_V/P_P) \tag{521.1}$$

$$\simeq kT\log_e(N_V/N_{AA}). \quad (2) \tag{521.1A}$$

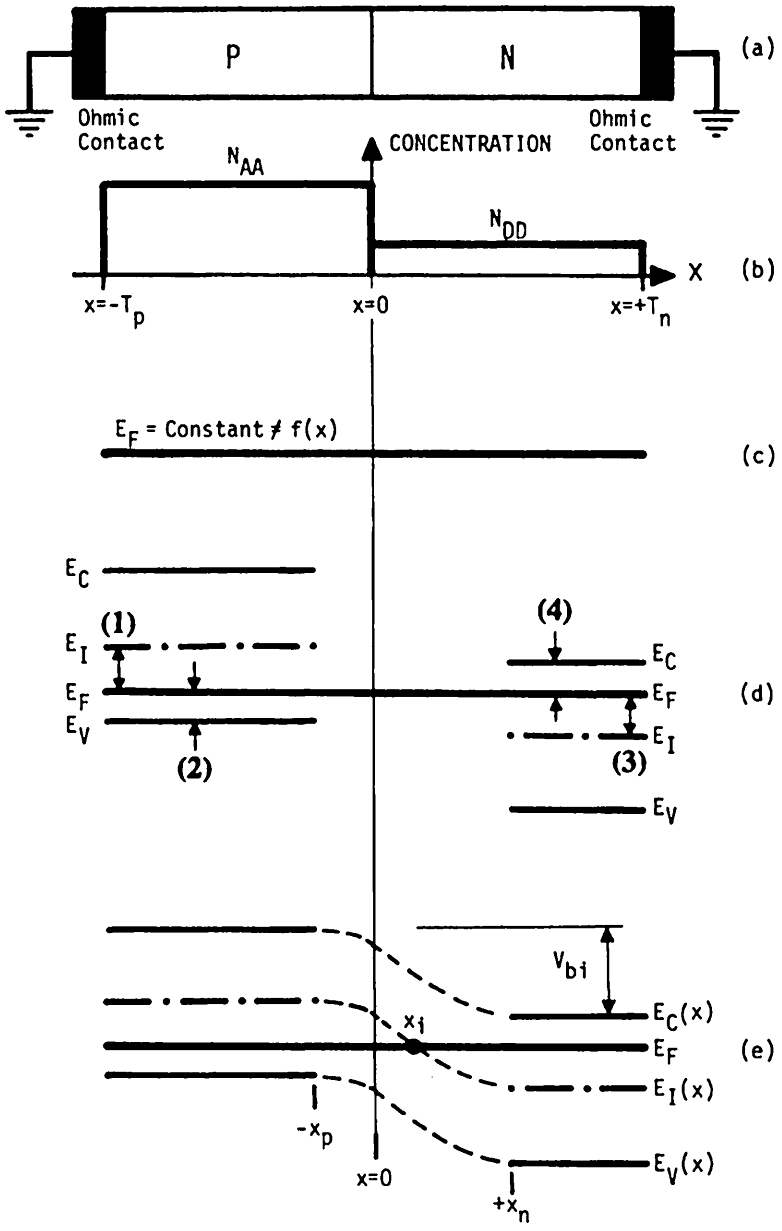

P
N
(a)
Ohmic Contact
Ohmic Contact
CONCENTRATION
N_AA
N_DD
X
(b)
x=-T_p
x=0
x=+T_n
E_F = Constant ≠ f(x)
(c)
E_C
(1)
E_I
E_F
E_V
(2)
(4)
E_C
E_F
E_I
(3)
E_V
(d)
V_bi
E_C(x)
x_i
E_F
(e)
E_I(x)
-x_p
x=0
E_V(x)
+x_n

Fig.521.1(a)-(e) The energy band diagram of the p-region and the n-region far away from the p/n junction boundary.

Following the same procedure just given for the p-type bulk, we use the Boltzmann approximation for the n-type Si bulk, (233.8), by assuming the n-type is not degenerate ($N < 10^{18}\mathrm{cm}^{-3}$). Then, the position of the Fermi energy level relative to the conduction band edge in the n-type region far away from the p/n junction, such as the back surface of the Si slice $x=+T_n$, is given by [labeled (4) in Fig.521.1(d)]

$$N = N_N = N_C \exp[-(E_C-E_F)/kT] \simeq N_{DD}$$

and

$$E_C - E_F \text{ (at } x=+T_n) = kT\log_e(N_C/N_N) \qquad (521.2)$$

$$\simeq kT\log_e(N_C/N_{DD}). \qquad (4) \qquad (521.2A)$$

Here N_N is the equilibrium electron concentration in an n-type bulk region which is nondegenerate or $N_N < \sim 10^{18}\mathrm{cm}^{-3}$.

Let us plug in some numbers. Consider Si at room temperature (296.57K), we have $N_C=2.75\times10^{19}$ and $N_V=1.3\times10^{19}$ cm^{-3} from (233.8B) and (233.10B). Let us take the case of $N_{AA}=10^{18}$ and $N_{DD}=10^{14}$ cm^{-3}, where N_{AA} is 10^{-2} of C_o used for the diffused p/n junction described in section 513 to allow the use of Boltzmann approximation. Then, the Fermi level positions labeled (2) and (4) in Fig.521.1(d) can be numerically computed from (521.1A) and (521.2A):

$$E_F-E_V \text{ (at } x=-T_p) = kT\log_e(1.3\times10^{19}/10^{18}) = 25.5544\mathrm{meV}\times2.5649$$

$$= 65.545\mathrm{meV} \qquad (2) \qquad (521.3)$$

and

$$E_C-E_F \text{ (at } x=+T_n) = kT\log_e(2.8\times10^{19}/10^{14}) = 25.55\mathrm{meV}\times12.54$$

$$= 320.46\mathrm{meV}. \qquad (4) \qquad (521.4)$$

The energy gap of Si has been measured very accurately over a wide range of temperature (<4K to >400K) and is $E_G=1.0908\mathrm{eV}$ at 296.57K from the empirical formula (512.2). Thus, we can also find the position of the Fermi level relative to the far or the minority carrier edge of the energy gap, i.e. E_C in the p-type bulk and E_V in the n-type bulk. These are

$$E_C-E_F \text{ (at } x=-T_p) = E_G - (E_F-E_V) = 1090.8 - 65.55$$

$$= 1025.25\mathrm{meV} \qquad (2) \qquad (521.5)$$

and

$$E_F-E_V \text{ (at } x=+T_n) = E_G - (E_C-E_F) = 1090.8 - 320.46$$

$$= 770.34\mathrm{meV}. \qquad (4) \qquad (521.6)$$

Intrinsic Fermi Level and Electric Potential

In the above analysis, we have computed the Fermi energy relative to the energy band edges, E_C and E_V, at the two ends of a p/n junction. However, the electric potential is commonly computed from the intrinsic Fermi level. The reason is that when the intrinsic Fermi level is used as the potential, the results of a device analysis are symmetrical with respect to electrons and holes. This symmetry considerably simplifies the analysis, reducing the algebra by nearly 50%. It also makes the result much easier to understand physically and to picture mentally.

The position of the intrinsic Fermi level, E_I, was given by (241.2) which shows that it is located nearly at the middle of the energy gap. More precisely, it is slightly below the midgap position in Si because in Si, the effective density of state of the conduction band, N_C, is slightly larger than that of the valence band, N_V. At 300K, we have

$$\begin{aligned} E_I &= (1/2)(E_C+E_V) + (3/4)kT\log_e(N_V/N_C) \\ &= (1/2)(E_C+E_V) + (3/4)kT\log_e(1.3\text{x}10^{19}/2.75\text{x}10^{19}) \\ &= (1/2)(E_C+E_V) - (0.75/2)kT \\ &= (1/2)(E_C+E_V) - 9.57\text{ meV}. \end{aligned} \tag{521.7}$$

Thus, E_I is about 10 meV below the midgap in Si. This is the reason that the intrinsic Fermi level is frequently referred to but incorrectly assumed to be exactly equal to the midgap position. It is incorrect since $\exp(qV/kT) = \exp(9.57/25.85) = 1.448$ which will give 44.8% error in carrier concentration calculations when assuming $E_I = E_{mg}$ (or E_{MG}) which would give 1.000.

From the numerical result of the position of the intrinsic Fermi potential or the intrinsic electric potential, we can also calculate the position of the Fermi level relative to the intrinsic Fermi level. These are labeled as (1) and (3) for the p-type and n-type end of the p/n junction in Fig.521.1(d). Using (521.3), (521.4) and (521.7), then

$$\begin{aligned} E_I - E_F \text{ (at } x=-T_p) &= E_I-E_V-(E_F-E_V) = (1/2)(E_C-E_V)-9.57-65.55 \\ &= (E_G/2)-9.57-65.55 = (1/2)\text{x}1090.8-9.57-66.55 \\ &= 470.28\text{ meV.} \qquad (1) \end{aligned} \tag{521.8}$$

$$\begin{aligned} E_F - E_I \text{ (at } x=+T_n) &= 1090.8/2 + 9.57 - 320.52 \\ &= 234.45\text{ meV.} \qquad (3) \end{aligned} \tag{521.9}$$

There is also a simpler way to compute the position of the Fermi energy relative to the intrinsic Fermi level. This is to use the expression of the electron and hole concentrations written in terms of the intrinsic carrier concentration, n_i, and the intrinsic Fermi level, E_I, given by (242.3) and (242.4). These equations are repeated here.

$$N = n_i \exp[(E_F-E_I)/kT] \tag{521.10}$$

$$P = n_i \exp[(E_I-E_F)/kT] \tag{521.11}$$

Note the symmetry of these two equations. N/n_i is just the reciprocal of P/n_i.

If our numerical calculations given in (521.8) and (521.9) are accurate, then we should be able to verify these results using (521.10) and (521.11). At 296.54K, we have $E_G = 1090.8$ meV which we have used above and $n_i = 1.000 \times 10^{10}$ cm^{-3}. Thus, from (521.11), the energy level positions labeled (1) and (3) in Fig.521.1(d) are respectively

$$E_I-E_F \text{ (at } x=-T_p) = kT\log_e(P_P/n_i) = 25.551\log_e(10^{18}/10^{10})$$

$$= 470.74 \text{ meV} \quad (1) \tag{521.12}$$

and

$$E_F-E_I \text{ (at } x=+T_n) = kT\log_e(N_N/n_i) = 25.551\log_e(10^{14}/10^{10})$$

$$= 235.37 \text{ meV} \quad (3) \tag{521.13}$$

These numbers are in excellent agreement with those computed in Eqs.(521.8) and (521.9). The small discrepancies come from calculator errors since they should give identical numerical results to all significant figures. Notice the much simpler calculation steps in (521.10) to (521.13) from using the intrinsic Fermi level, intrinsic carrier concentration, and the carrier concentrations relationships given by (521.10) and (521.11).

522 Equilibrium Potential Barrier Height in a p/n Junction

The energy band diagram in Fig.521.1(d) shows that the position of E_C (or E_I or E_V) on the p-type side of the junction is higher than that on the n-type side. Thus, there is a potential energy step going from the n-type side to the p-type side. This is known as the **equilibrium potential barrier height** of a p/n junction or for short, the **equilibrium barrier height**. It is commonly known as the **built-in potential barrier**, also used in MOSC. In potential unit (Volt), it is denoted by the symbol V_{bi}, V_{BI}, V_{PNbi} or $V_{PN\text{-}bi}$ from p/n junction built-in potential. It cannot be measured by connecting a voltmeter to the two terminals of a p/n junction diode since it is a equilibrium property. In order to measure V_{bi} or cause the voltmeter to deflect, a charged particle current is needed, such as the mobile ion current in a battery which causes a electron current to flow in the conductor wires connected to the voltmeter. Fortunately, V_{bi} can be measured indirectly by extrapolation of a nonequilibrium characteristic, such as the capacitance-voltage or current-voltage curve of a p/n junction diode. The fundamental esoteric point is that it is the gradient of the quasi-Fermi potential that drives the diffusion plus the drift current as shown by (331.1) to (331.4A) while the gradient of the electrostatic potential produces only a electric or electrostatic field that drives only a drift current. This drift current is completely canceled by the diffusion current at equilibrium and hence the potential that drives the drift current cannot be measured.

From the calculations we just made of the position of the Fermi level relative to the energy band edges, we can compute the potential barrier height, which is explicitly labeled in Fig.521.1(e). From Fig.521.1(d) we see that it is given by $E_C(-T_p)-E_C(+T_n)$ or $E_I(-T_p)-E_I(+T_n)$ or $E_V(-T_p)-E_V(+T_n)$. Using the the intrinsic Fermi level and (521.12) and (521.13), then

$$+\,qV_{bi} = E_I(-T_p) - E_I(+T_n) = [E_I(-T_p) - E_F] + [E_F - E_I(+T_n)]$$

$$= kT\cdot\log_e(P_P/n_i) + kT\cdot\log_e(N_N/n_i) \qquad (522.1)$$

or

$$V_{bi} = (kT/q)\cdot\log_e(P_P N_N/n_i^2) \quad \text{(Volt)} \qquad (522.2)$$

$$\simeq (kT/q)\cdot\log_e(N_{AA} N_{DD}/n_i^2) \quad \text{(Volt)} \qquad (522.2A)$$

Using the numerical values in the illustration just given, $N_{AA}=10^{18}$, $N_{DD}=10^{14}$ and $n_i=10^{10}$ cm^{-3} at room temperature (296.57K), then

$$V_{bi} = 25.5554\cdot\log_e(10^{18}\cdot 10^{14}/10^{20}) = 706.12\text{mV} \qquad (522.3)$$

This can be verified by adding the numerical solutions of (521.12) and (521.13) to give 470.74+235.37=706.11 mV. Many significant figures must be retained in order to give an accurate electron or hole concentration due to the exponential dependence, $\exp(V_{bi}/kT)$, since kT/q is only about 25 mV. For example, 1% error in V_{bi}, ± 7.06mV, would give $\exp(\pm 7.06/25)=1.33^{\pm 1}$ or 33% error in $\exp(V_{bi}/kT)$ or 33% error in the electron or hole concentrations.

Using (521.12) and (521.13) to calculate separately, we get the additional information of how the barrier is divided between the n-type and p-type side. The division point is the point labeled x_I in Fig.521.1(e). This is the point where the Fermi level intercepts the intrinsic Fermi level, $E_I(x)$. It is the point where the electron and hole concentrations are equal and equal to the intrinsic carrier concentration. It will be termed the **intrinsic boundary** and is the **electronic junction boundary**. Note it does not coincide with the metallurgical boundary, x_j, which is located at $x_j=0$ in the coordinate choice given in Fig.521.1(b).

There is another way of computing the potential barrier height which gives a very important additional physical insight on the existence of a potential barrier separating the n-type and p-type regions. At electronic equilibrium, the electron current and the hole current must **individually** be zero. However, the individual drift and diffusion component of the electron current (or the hole current) need not be zero. In an atomically uniform semiconductor, the diffusion and drift currents are indeed individually zero at equilibrium. But they are not zero in a p/n junction or a semiconductor containing a material nonuniformity that can give rise to an ion concentration gradient. For example, the electron diffusion current cannot be zero in a p/n junction structure because there are more electrons on the n-type side than on the p-type side, causing electrons to flow from the higher concentration n-type side to the lower concentration p-type side, by diffusion. However, this diffusive flow cannot persist forever because when the electrons move from the n-type side to the p-type side, the n-type side will have fewer electrons than its equilibrium value. This causes a net positive space charge to build up within the n-type side which is given by $\rho(x)=q(P-N+N_{DD})$. Furthermore, the electrons which have moved to the p-type side by diffusion will charge up the p-type side negatively, given a space-charge density of $\rho(x)=q(P-N-N_{AA})$.

This space charge will create an electric field which will produce an electron drift current in the direction opposite to that of the electron diffusion current. At electronic equilibrium, the net electron current is zero. Thus, the electron diffusion current from the n-side to the p-side is exactly balanced by the electron drift current from the p-side to the n-side.

A similar description can be made for the holes which gives an electric field in exactly the same direction and at exactly the same magnitude as that produced by the electrons!

The zero net electron current or zero net hole current can be used to compute the potential barrier height of a p/n junction at equilibrium. From (320.10) or (350.3) we have the following one-dimensional equilibrium solution,

$$0 = J_{Nx} = q\mu_n N E_x + qD_n(dN/dx). \qquad (522.4)$$

It is integrated using $E_x=-dV_I/dx$ and the Einstein relationship $D_n/\mu_n=kT/q$, to give

$$\int_1^2 (D_n/\mu_n)dN/N = (kT/q)\cdot\log_e(N_2/N_1)$$

$$= -\int_1^2 E_x dx = +\int_1^2 (dV_I/dx)dx = V_I(x_2) - V_I(x_1) \qquad (522.5)$$

where $N_1 = N(x_1)$ and $N_2 = N(x_2)$. x_1 and x_2 are two unspecified locations in the the semiconductor, so this equilibrium result is perfectly general and its only limitation is that it is derived for the one-dimensional case. It is general since we did not use any restrictions on how to select N_1 and N_2 or x_1 and x_2. A second limitation of (522.5) is that it is applicable only to low electron concentration regions so that the Boltzmann approximation can be used to approximate the Fermi-Dirac distribution function. The Boltzmann approximation was used in section 233 to give the electron and hole concentrations, (233.8) and (233.10). Thus, at equilibrium and in the low carrier concentration layers, we have

$$V_I(x_1) - V_I(x_2) = + (kT/q)\cdot\log_e[N(x_1)/N(x_2)] \qquad (522.6)$$

Now, let us apply this result to the calculation of the potential barrier height in a p/n junction at equilibrium. From Figs.521.1(b) and (c), we have

$$x_1 = -T_p$$

$$x_2 = +T_n,$$

$$N_1 = N(x_1) = N(-T_p) = N_P \simeq n_i^2/N_{AA}$$

and

$$N_2 = N(x_2) = N(+T_n) = N_N \simeq N_{DD}.$$

Thus, using these in (522.5), we have

$$V_D = V_I(x_2) - V_I(x_1)$$

$$= (kT/q)\cdot\log_e[N(x_2)/N(x_1)] = (kT/q)\cdot\log_e(N_N/N_P) \qquad (522.7)$$

$$\simeq (kT/q)\cdot\log_e(N_{DD}N_{AA}/n_i^2). \qquad (522.7A)$$

The approximate result, (522.7A), applies to the case when both sides or both locations, x_1 and x_2, are extrinsic while (522.7) applies even if one side is intrinsic. When both sides are intrinsic, we do not have a p/n junction nor a potential barrier. The results given by (522.7) and (522.7A) show that a potential difference appears between the p-side and n-side of a p/n junction due to the difference in concentration of electrons (or holes) on the two sides and the concentration difference causes diffusion. These results are identical to those obtained in (522.1) and (522.2) by a different method. The present derivation emphasizes diffusion and the condition of zero current for each charge carrier species based on an elaboration of the basic physics that the diffusion current is exactly balanced or cancelled by the drift current at equilibrium. Thus, the potential barrier height is also known as the diffusion barrier height, and V_D is known as the diffusion voltage or diffusion potential, hence the subscript D in V_D. It is identical to V_{bi} derived previously in (522.1) and (522.2). This term, diffusion potential, was coined by those who derived the result using the zero-current method just presented, emphasizing diffusion but somewhat overlooking the equally important drift current which must contribute equally in order to exactly balance the diffusion current. Thus, it is also the drift potential barrier height. But, indeed, diffusion is the cause arising from a carrier concentration gradient owing to a built-in dopant impurity concentration gradient and drift is the effect. A most correct term, unfortunately devoid of any physics, is 'built-in potential' using the symbol V_{BI} or V_{bi}. Some authors have also used the symbol, V_B, where B=barrier, instead of V_D or V_{bi}. This is conceptually defective since the potential barrier height depends also on the applied voltage while the diffusion potential, V_D, is an intrinsic property of the material independent of the applied voltage. To avoid this confusion and a potential mix-up of V_{BI} with V_B, or a mix-up of the diffusion potential, V_D, with applied drain voltage, also V_D, in the MOS transistor, we use V_{bi} or V_{PNbi}, V_{EBbi}, V_{CBbi}, and V_{Dbi} for multi-junction devices. We use V_D occasionally when a confusion will not occur.

The foregoing discussion on the notation difficulties is presented to illustrate the underlying device physics of the p/n junction and to emphasize the importance of selecting proper symbols in order to avoid misleading, incomplete, confusing, or incorrect physical concepts.

523 Equilibrium Potential Variation in a p/n Junction

Thus far, some properties of the equilibrium energy band diagram of a p/n junction are obtained without solving any complicated algebraic and differential equations. The result shows that there is a potential energy step or potential barrier of V_{bi} (or V_D) which the electron must climb while going from the n-type side to the p-type side of a p/n junction.

The procedure thus far allows us only to draw three dotted curves in Fig.521.1.1(e) to approximate the potential variation from the p-side to the n-side. To obtain this potential variation quantitatively and accurately, we must solve the Poisson Equation which is the fifth equation, (350.5), of the six Shockley equations listed in section 350. It can be solved analytically or by numerical integration on a computer at equilibrium without having to also solve the other five Shockley equations. The reason is that there is no current so that the other five equations all give null result. It is a special property at equilibrium. We need to find the exact or accurate analytical solution at equilibrium since it gives us an accurate reference point from which to extend the analysis to nonequilibrium situations created by an applied force, such as a terminal voltage, an electric and magnetic field, an exposure to light, heat, temperature gradient, and a mechanical force.

In one-dimensional form, the Poisson equation given by (350.5) simplifies to

$$d(\varepsilon_s E/dx) = - \varepsilon d^2V_I/dx^2 = q(P - N + N_{ION}) \qquad (523.1)$$

where $N_{ION} \equiv N_{AA}$ or N_{DD}, assuming all the dopant acceptor and donor impurity atoms are ionized. We also assume very low trap density so that $N_{TT} << N_{AA}$ and $N_{TT} << N_{DD}$. Using the Boltzmann approximations for the electron and hole concentrations at thermal equilibrium given by (521.10) and (521.11),

$$N = n_i\exp[(E_F-E_I)/kT] = n_i\exp[-q(V_F-V_I)/kT] \qquad (523.2)$$

and

$$P = n_i\exp[(E_I-E_F)/kT] = n_i\exp[+q(V_F-V_I)/kT] \qquad (523.3)$$

and $E = -dV_I/dx$, then, the Poisson equation, (523.1), becomes

$$-\varepsilon_s d^2V_I/dx^2 = q\{n_i\exp[q(V_F-V_I)/kT] - n_i\exp[-q(V_F-V_I)/kT] + N_{ION}\} \qquad (523.4)$$

which is known as the **Poisson-Boltzmann equation** or the **Poisson equation in the Boltzmann approximation.** Evidently it is nonlinear because of the two exponential terms from carrier concentration. For degenerate or heavily doped semiconductor and metal, the electron (hole) concentration is so high ($>10^{18}$cm^{-3}) that the Boltzmann approximation is no longer valid. Then, the Fermi-Dirac integral must be used for the electron and the hole concentrations in place of the two exponentials which is then the **Poisson-Fermi equation.**

The Poisson-Boltzmann equation given by (523.4) can be integrated analytically only once provided that N_{AA} and N_{DD} are spatially constant. It cannot be integrated analytically at all if one of N_{AA} and N_{DD} is not spatially constant. Thus, it does not help us to understand the basic properties if we try to solve it exactly since the second integration must be made numerically which obscures the basic physics. Instead, we shall obtain an explicit analytical but approximate solution to help us to understand the mathematics and its underlying semiconductor device physics. To this aim, we note that the Poisson-Boltzmann equation, (523.4), can be simplified if we neglect the electron and hole concentration terms, P and N, since then N_{ION} ($\equiv N_{AA}$ on the p-side and $\equiv N_{DD}$ on the n-side) is the only remaining charge density term and it is a function of the independent variable, x, and not the dependent variable, $V_I(x)$. This is known as the **depletion approximation** from the assumption of carrier depletion, $N=0$ and $P=0$ or more precisely, $|N-P| << |N_{ION}|$. The term, **depletion,** was introduced by Shockley in order to obtain analytical solutions when he invented the bipolar junction transistor in 1949. The Poisson equation (523.4) then becomes

$$-\ \varepsilon_s d^2V_I/dx^2 = qN_{ION}(x) \quad \textbf{(carrier depletion approximation)} \qquad (523.4A)$$

It is a particularly good approximation if a negative voltage is applied to the p-side contact relative to the n-side contact of a p/n junction because then holes will be drawn back towards the p-side contact and electrons will be drawn back towards the n-side contact by the applied voltage, leaving a region adjacent to the p/n boundary where electrons and holes are depleted.

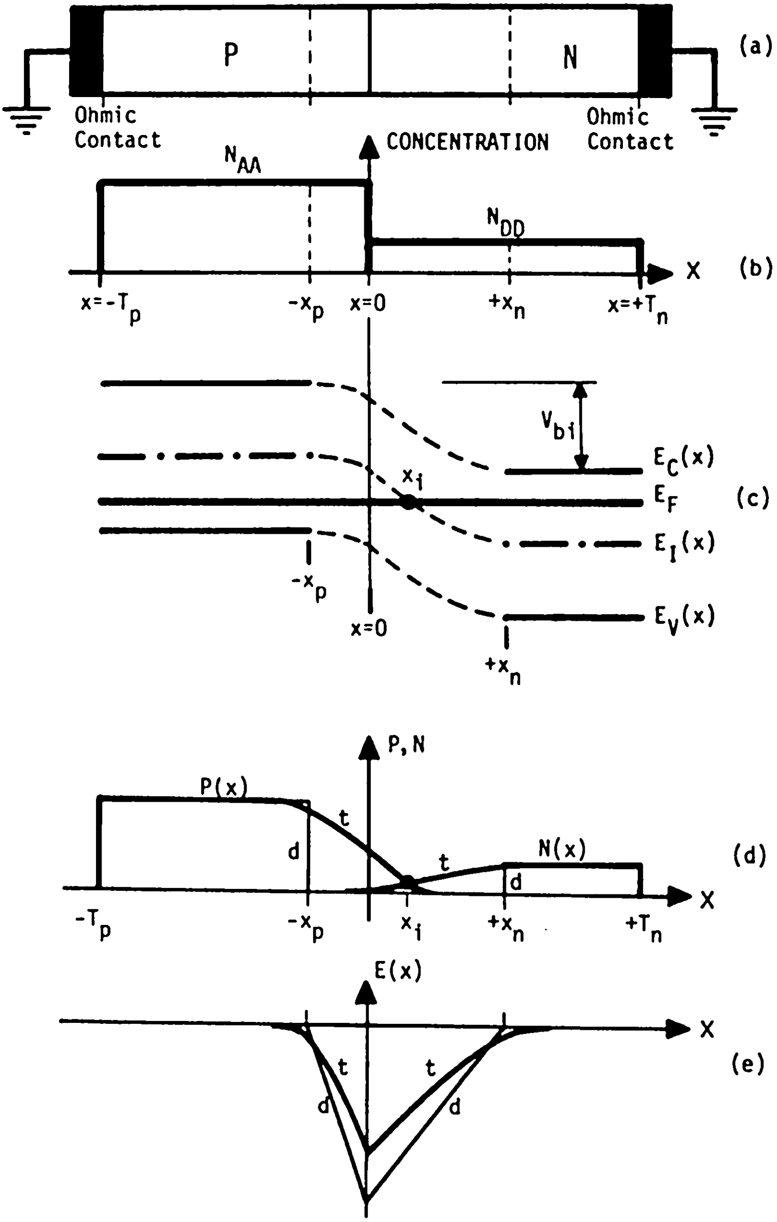

P
N
(a)
Ohmic Contact
Ohmic Contact
CONCENTRATION
N_{AA}
N_{DD}
X
(b)
$x=-T_p$
$-x_p$
$x=0$
$+x_n$
$x=+T_n$
V_{bi}
$E_C(x)$
x_i
E_F
(c)
$E_I(x)$
$-x_p$
$x=0$
$E_V(x)$
$+x_n$
P,N
P(x)
t
d
N(x)
t
d
(d)
X
$-T_p$
$-x_p$
x_i
$+x_n$
$+T_n$
E(x)
X
(e)
t
t
d
d

Fig.523.1(a)-(e)) The transition layer of the p/n junction boundary at thermal equilibrium. (a) The cross sectional view. (b) The ionized impurity concentration profile. (c) The potential energy band diagram. (d) The electron and hole distributions. (e) The electric field variation.

In the depletion approximation given by (523.4A), we can even lift the spatial constancy restriction on N_{AA} and N_{DD} and still obtain analytical solutions from the Poisson equation. It is evident that such analytical solutions are obtainable by double integration of (523.4A) using the appropriate boundary conditions for $V_I(x)$ and $dV_I(x)/dx|$ at $x=x_{boundary}$. The reason is that $N_{AA}=N_{AA}(x)$ and $N_{DD}=N_{DD}(x)$ and are not functions of the unknown $V(x)$. Thus, $\int\int N_{AA}(x)dxdx$ and $\int\int N_{DD}(x)dxdx$ can be explicitly evaluated in terms of known or tabulated functions. The endeavor of solving the depletion Poisson equation for p/n junctions of all kinds of dopant impurity distributions, $N_{AA}(x)$ and $N_{DD}(x)$, has occupied many device engineers in the past four decays since Shockley wrote the first paper on the electrical properties of p/n junction and p/n (bipolar) junction transistor in 1949. To introduce and emphasize the basic junction physics, we shall next obtain the simplest solution of the depletion model, that is, N_{AA} and N_{DD} are independent of position but not equal.

The energy band diagram shown in Fig.521.1(e) is redrawn in Fig.523.1(c). The layer where the potential drop occurs in the vicinity across the p/n junction boundary is known as the **space-charge layer**. It is also known as the **transition layer** through which the conductivity type changes from p-type to n-type. In this figure, it is extended from $-x_p$ to $+x_n$ and encloses the 'metallurgical' junction or **dopant impurity boundary** denoted by the plane $x=0$. This differs from the **electronic boundary**, x_I, where the conductivity changes from p-type to n-type. In the **depletion approximation**, electron and hole concentrations are assumed to be zero in the space-charge layer, so that the transition layer is also known as the **depletion layer**. However, space-charge layer is the more appropriate term to use since the term 'depletion layer' comes from the assumption of $P=N=0$ which may not be valid under general conditions. In the literature as well as most textbooks, one will find that this distinction is not made, consequently causing considerable misconception when the term depletion layer is used. So the correct description is **'the solutions of the potential in the space-charge layer using the depletion approximation'** instead of 'the solutions of the potential in the depletion layer'.

The boundary conditions and differential equations in the depletion approximation are

$$\rho = 0 \quad \text{and} \quad E_x = 0 \qquad -T_p<x<-x_p \qquad (523.5)$$

$$\rho = -qN_{AA} = -\varepsilon_s d^2V/dx^2 \qquad -x_p<x<0 \qquad (523.6)$$

$$\rho = +qN_{DD} = -\varepsilon_s d^2V/dx^2 \qquad 0<x<+x_n \qquad (523.7)$$

$$\rho = 0 \quad \text{and} \quad E_x = 0 \qquad +x_n<x<T_n \qquad (523.8)$$

$$dV/dx = 0 \qquad x=-x_p \text{ and } x=x_n \qquad (523.9)$$

$$V(x) = 0 \qquad x=0 \qquad (523.10)$$

and

$$V(+x_n) - V(-x_p) = V_D \equiv V_{bi}. \qquad (523.11)$$

Equations (523.5) and (523.8) are the charge neutrality condition we previously assumed for the regions far away from the p/n boundary when we computed the diffusion or built-in potential, V_{bi}. We have moved the neutrality condition inwards toward the p/n boundary, from $+T_n$ to $+x_n$ and from $-T_p$ to $-x_p$. This move is an approximation known as the **'quasi-neutrality approximation'**. 'Quasi' means that it is not exactly but nearly, first used and defined by Shockley in 1949 for solid-state electronics. It assumes that the space-charge neutrality is approximately attained in the two layers, $-T_p$ to $-x_p$ and x_n to T_n. A more concise and quantitative criterion of quasi-neutrality is that the potential change in a quasi-neutral layer is smaller than a thermal voltage drop across a Debye length, kT/qL_D. This slow change corresponds to a small electric field and it can be derived by an iterative analytical approximation using the space-charge neutral results to be obtained. The quasi-neutrality assumption is reasonable since we expect the influences of the p/n transition to diminish at large distances from the p/n boundary and we assume that the zero influence begins at $x=-x_p$ and $x=+x_n$ from the p/n boundary at $x=0$.

Equations (523.6) and (523.7) are the depletion approximations in the space-charge layer which we previously discussed. The $dV/dx=0$ boundary condition given by (523.9) and the $E_x=0$ condition in the quasi-neutral regions given by (523.5) and (523.8) are approximate results at zero current owing to assumed constant impurity concentrations. The $V(x)=0$ boundary condition given by (523.10) is exact and is an arbitrary but mathematically convenient choice, since it gives solutions with fewer terms. Equation (523.11) is the total potential drop equaling the diffusion potential or built-in potential obtained previously.

The depletion approximation is shown in Figs.523.1(d) and (e). They are the light curves (or lines) labeled with the letter d. The heavy curves (labeled t for true) are the exact variations of the electron and hole concentrations, N(x) and P(x), and the electric field, E(x). In the depletion approximation, the electron and hole concentrations are assumed to be zero inside the space-charge layer, $-x_p \le x \le x_n$ and the resultant electric field is linearly dependent on position which is obtained in the following analysis.

The Poisson equation on the two sides of the p/n boundary, given by (523.6) and (523.7), can be analytically integrated twice using the boundary conditions given by (523.9), (523.10) and (523.11). The results are

p-type side $(-x_p \le x \le 0)$ n-type side $(0 \le x \le +x_n)$

$$dV/dx=(qN_{AA}/\varepsilon_s)(x+x_p) \qquad dV/dx=(qN_{DD}/\varepsilon_s)(x_n-x) \tag{523.12}$$

$$V(x)=(qN_{AA}/2\varepsilon_s)[(x+x_p)^2-x_p^2] \qquad V(x)=(qN_{DD}/2\varepsilon_s)[x_n^2-(x_n-x)^2]. \tag{523.13}$$

These results may be verified by checking to determine if they satisfy the boundary conditions given by (523.9) and (523.10). They show that the electric field, given by $E(x)=-dV(x)/dx$, is linearly dependent on position. The linear dependencies are shown in Fig.523.1(e) by the light lines labeled d.

These results also show that the electric potential versus position is parabolic on each side of the p/n boundary, and is smoothly joined at the p/n boundary, $x=0$ where the maximum slope or maximum electric field occurs. These parabolic dependences are shown by the light dash curves in Fig.523.1(c) and are known as the **parabolic potential variations**.

We can also get an explicit relationship between the total space-charge layer thickness and the built-in potential. First substitute (523.13) into (523.11) to give

$$V_{bi} \equiv V_D = V(x_n) - V(-x_p) = (q/2\varepsilon)[N_{DD}x_n^2 + N_{AA}x_p^2]. \qquad (523.14)$$

Next, we use the condition that the electric field at the p/n boundary, $x=0$, must be continuous. This continuity condition may not be true generally if a thin or atomic layer of trapped charges exists on the p/n boundary, for example, the charged interface traps that can exist at the semiconductor heterojunction interface between two dissimilar semiconductors or at the oxide/semiconductor interface in an MOS capacitor. Applying this to dV/dx given by (523.12) at $x=0$, we have

$$dV(0)/dx = (q/\varepsilon_s)N_{AA}x_p = (q/\varepsilon_s)N_{DD}x_n$$

or

$$N_{AA}x_p = N_{DD}x_n. \qquad (523.15)$$

Let us denote the total space-charge layer thickness by x_{pn}, that is

$$x_{pn} = x_n + x_p \qquad (523.16)$$

then

$$x_p = x_{pn}[N_{DD}/(N_{AA}+N_{DD})] \qquad (523.17A)$$

and

$$x_n = x_{pn}[N_{AA}/(N_{AA}+N_{DD})]. \qquad (523.17B)$$

These two results are put in a form designed to be remembered easily: the higher doped side has a smaller thickness, and the total density of the impurity on each side is equal, $x_nN_{DD}=x_pN_{AA}$, which follows from the fact that the total charge contained in the space-charge layer, $-x_p<0<+x_n$, must vanish. Substituting these into (523.14), we get

$$V_{bi} = (q/2\varepsilon_s)x_{pn}^2[N_{AA}N_{DD}/(N_{AA}+N_{DD})] \qquad (523.18)$$

or

$$x_{pn} = \sqrt{2\varepsilon_s V_{bi}/qN_M} \qquad (523.19)$$

where N_M is the geometric average or mean of N_{AA} and N_{DD} defined by

$$N_M = N_{MEAN} = N_{AA}N_{DD}/(N_{AA}+N_{DD})$$

or

$$1/N_M = 1/N_{AA} + 1/N_{DD}. \qquad (523.20)$$

The maximum electric field occurs at the metallurgical boundary of the junction, $x=0$, and is given by (523.12). Using x_p or x_n from (523.17) and x_{pn} from (523.19), the maximum electric field is given by

$$E_{MAX} = |dV(0)/dx| = \sqrt{2qV_{bi}N_M/\varepsilon_s} \qquad (523.21)$$

$$= 2V_{bi}/x_{pn} \qquad (523.22)$$

which shows that the maximum electric field is twice the average electric field given by V_{bi}/x_{pn}. The larger maximum electric field is due to the linear variation of the electric field with distance in the space-charge layer shown by (523.12). It is highly dependent on the dopant impurity profile. For a **linearly-graded** junction with N_{DD}-N_{AA} = ax where a is the concentration gradient, the maximum electric field is $3V_{bi}/2x_{pn} = 1.5V_{bi}/x_{pn}$.

The result given by (523.18) is particularly easy to derive if one cannot remember which factors are in the numerator and denominator. All one has to do is to look at the Poisson equation: $\varepsilon_s d^2V/dx^2 = qN_M$ and integrate it twice with all constants of integration set to zero. Then, one gets a result which will lead to (523.19) after setting $x=x_{pn}$. This is really a dimension check but it is also a quick way to get to the correct final expression given by (523.19).

The pair of solutions, (523.19) and (522.7A), listed below again for convenience, give the solution of the space-charge layer at equilibrium for a given set of numerical values for N_{DD}, N_{AA}, T, n_i and ε_s.

$$V_{bi} = (kT/q)\cdot\log_e(N_{DD}N_{AA}/n_i^2) \qquad \text{from (522.7A)} \qquad (523.23A)$$

$$x_{pn} = \sqrt{2\varepsilon_s V_{bi}/qN_M} \qquad \text{from (523.19)} \qquad (523.23B)$$

where $N_M = N_{AA}N_{DD}/(N_{AA}+N_{DD})$. In addition, the space-charge layer thicknesses on two sides of the p/n boundary are

$$x_p = x_{pn}N_{DD}/(N_{AA}+N_{DD})$$

and

$$x_n = x_{pn}N_{AA}/(N_{AA}+N_{DD}). \qquad (523.23C)$$

Let us plug in the numbers used previously: $N_{AA}=10^{18}$, $N_{DD}=10^{14}$, T=296.6K, $n_i=10^{10}$ and $\varepsilon_s=11.68\text{x}8.854\text{x}10^{-14}$ to get a feel of the order of magnitude of the thickness of the layers. With these dopant densities, we had V_{bi}=706.12mV in (522.3), then the space-charge layer thickness is

$$x_{pn} = \sqrt{2\text{x}11.68\text{x}8.854\text{x}10^{-14}\text{x}0.70612/1.6\text{x}10^{-19}\text{x}10^{18}\text{x}10^{14}(10^{18}+10^{14})^{-1}}$$

$$\simeq \sqrt{2\text{x}11.68\text{x}8.854\text{x}10^{-14}\text{x}0.70612/1.6\text{x}10^{-19}\text{x}10^{14}} \qquad (523.24)$$

$$= 2.77\text{x}10^{-4}\ \text{cm} = 2.77\ \text{microns.} \qquad (523.25)$$

Note, the approximation made in (523.24) is $N_M=N_{AA}N_{DD}/(N_{AA}+N_{DD})\simeq N_{AA}$ (the smaller one of N_{AA} and N_{DD}).

We may also compute the thicknesses of the space-charge layer on each side of the p/n junction boundary using (523.23C) and the above results to get

$$\begin{aligned} x_p &= x_{pn}N_{DD}/(N_{DD}+N_{AA}) \\ &= 2.77\text{x}10^{14}/(10^{18}+10^{14}) = 2.77\text{x}10^{-4}\ \text{microns} \qquad (523.26) \\ x_n &= x_{pn}N_{AA}/(N_{DD}+N_{AA}) \\ &= 2.77\text{x}10^{18}/(10^{18}+10^{14}) = 2.77\ \ \text{microns.} \qquad (523.27) \end{aligned}$$

These results show that the space-charge layer is almost entirely on the p-type side, which is the lowly doped side ($x_p=2.77$ microns $\simeq x_{pn}$). The space-charge layer thickness on the n-type side of the boundary is 10,000 times smaller than the total space-charge layer as indicated by (523.26) due to the 10,000 higher dopant impurity concentration on the n-type side than the p-type side.

Whenever we have a large (a factor of ten or more) difference in the acceptor and donor dopant impurity concentration on the opposite sides of the p/n junction boundary, we have the above result: the space-charge layer is mainly on the lower doped side of the p/n boundary. Such a p/n junction is a highly **asymmetrical p/n junction**. In contrast, a junction with $N_{AA}=N_{DD}$ (a rarity in practice) is known as a **symmetrical junction**.

A highly asymmetrical p/n junction is commonly known as a p+/n or p^+/n junction ($N_{AA}>>N_{DD}$) and n+/p or n^+/p junction ($N_{DD}>>N_{AA}$). The '+' sign denotes **high** dopant impurity concentration.

The **slash** / notation convention I introduced and have used to denote a junction boundary since the 1970's, such as n+/p or n^+/p and p+/n or p^+/n, has gradually been adopted by transistor authors since the four-decade old tradition and habit of using a dash, -, are hard to change. My notation was selected and designed to show that there is a boundary or interface between the n+ side and the p side. The major reason for using '/' is to avoid the confusion from the traditional usage of a dash sign '-'. Consider the traditional usage, p^+-n or p+-n, very confusing indeed. Still more confusion if one has a lowly doped n-type layer denoted by n-, then one would have to write p+-n- which is obscurely confusing. Our notation would express this junction clearly by p+/n- . This slash notation can be further extended if we also want to explicitly indicate the ohmic contact metals. For example if we use aluminum on the front and gold on the back surfaces for ohmic contacts, we would have Al/p^+/n/Au, a very simple and telling notation. A further example of a structure encountered often in a VLSI chip is Al/n-poly-Si/O/p-Si/n^+Si/Al in the p-well of a CMOS inverter. The students can replace / by - and reach the same conclusion that the choice of / is aesthetically more appealing.

In summary, for an abrupt p/n junction in the depletion approximation, we have the following key results.

(1) Zero mobile carrier space charge in the space-charge layer, $N(x)=P(x)=0$.

(2) Electric field varies linearly with position in the space-charge layer. The maximum electric field occurs at the metallurgical junction, $x=0$, and it has a magnitude twice as large as the average.

(3) Electric potential varies parabolically with position in the space-charge layer.

(4) In a p+/n junction, the space-charge layer is mainly on the lower doped side of the p^+/n boundary, which is the n-side in this case.

(5) We have a simple way to relate the diffusion or built-in potential barrier height to the thickness of the space-charge layer. This is the quick integration of the Poisson equation twice by disregarding the boundary conditions. It helps to remember what factors are in the numerator and denominator but one must remember what the Poisson equation looks like.

(6) I have introduced a new notation to describe the makeup of a p/n junction diode and a multi-junction device, by using the slash '/' to indicate the junction boundaries or interfaces, such as Al/p^+/n/Au.

524 Application of the Gauss Theorem to a p/n Junction

The thickness of the space-charge layer on the opposite side of the p/n junction boundary, given by (523.17), can be readily derived by the Gauss Theorem as a short cut. The reason is that the Gauss Theorem is itself derived by integrating the Poisson equation. Gauss Theorem states that the net space-charge density **inside a volume enclosed by a conducting surface must be zero.** In the one-dimensional p/n junction example, the two conducting surfaces are the metal layers at the two surfaces, $x=-T_p$ and $x=+T_n$, of the semiconductor slice. The enclosed volume is the semiconductor slice between the two parallel planes, located at $x=-T_p$ and $x=+T_n$ as illustrated by the dashed 'pill-box' in Fig.524.1(a).

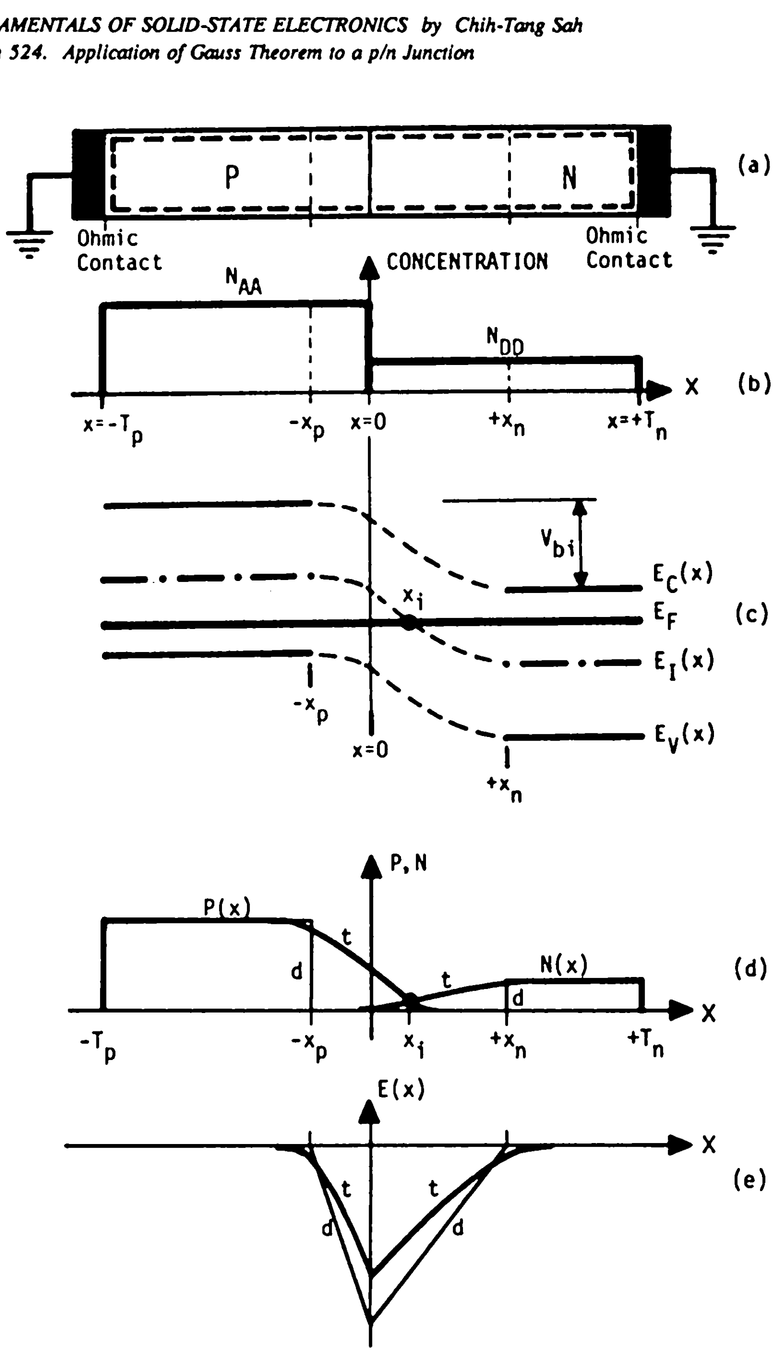

(a)
P
N
Ohmic Contact
Ohmic Contact
CONCENTRATION
N_{AA}
N_{DD}
X
(b)
$x=-T_p$
$-x_p$
$x=0$
$+x_n$
$x=+T_n$
V_{bi}
$E_C(x)$
x_i
E_F
(c)
$E_I(x)$
$-x_p$
$E_V(x)$
$x=0$
$+x_n$
P, N
P(x)
t
d
N(x)
t
d
(d)
X
$-T_p$
$-x_p$
x_i
$+x_n$
$+T_n$
E(x)
X
(e)
t
t
d
d

Fig.524.1 Application of Gauss Theorem to p/n junction. (a)-(e) are from Figs.523.1(a)-(e). Gauss theorem applied to the metallic 'pill box' in dashed lines in (a).

This one-dimensional Gauss Theorem can be readily derived by integrating the Poisson equation from $-T_p$ to $+T_n$ using the space-charge distributions given in Eqs.(523.5) to (523.4). We have

$$\int_{-T_p}^{T_n}\rho(x)dx = \int_{-T_p}^{-x_p}\rho(x)dx + \int_{-x_p}^{x_n}\rho(x)dx = \int_{T_p}^{T_n}\rho(x)dx = \int_{-x_p}^{x_n}\rho(x)dx \qquad (524.1)$$

since $\rho=0$ in the layers $-T_p < x < -x_p$ and $x_n < x < T_n$. The last integral can be integrated explicitly across the p/n junction space-charge layer to give

$$\int_{-x_p}^{x_n}\rho(x)dx = \int_{-x_p}^{x_n}\varepsilon_s(dE/dx)dx = \int_{-x_p}^{x_n}\varepsilon_s dE = E(x_n)-E(x_p) = 0-0 = 0. \qquad (524.2)$$

This is the Gauss Theorem in our one-dimensional p/n junction diode where we assume that the space charge and the electric field are zero outside of the p/n junction space-charge layer and at the two surfaces of the space-charge layer, $x=-x_p$ and $x=+x_n$. We now evaluate the integral of the space-charge density to get the total space-charge enclosed inside the space-charge layer and set it to zero in view of the Gauss Theorem. Thus,

$$0 = \int_{-x_p}^{x_n}\rho(x)dx = \int_{-x_p}^{0}-qN_{AA}dx + \int_{0}^{x_n}qN_{DD}dx = -qN_{AA}x_p + qN_{DD}x_n$$

or

$$N_{AA}x_p = N_{DD}x_n. \qquad (524.3)$$

This is identical to (523.15) which was obtained by integrating the Poisson Equation separately on each of the two parts of the space-charge layer rather than by integrating the space-charge density, $\rho(x)$, through the entire space-charge layer. We see in this example that we can use the Gauss Theorem to obtain results quickly. In order to use the Gauss Theorem, we must know the position of the demarcation boundary or zero-electric-field surface, beyond which the electric field and the space charge are zero.

525 Comparing the Depletion Approximation with the Exact Solution

The exact results of this problem are sketched based on physical intuition. They can be obtained quickly by numerical integration using a high speed computer or even a hand-held calculator such as the Hewlett-Packard model 11c. In Fig.524.1(d), the 'exact' or true electron and hole concentrations in the space-charge layer are represented by the heavy lines and curves and labeled t. They can be obtained from the Boltzmann relationships

$$P(x) = n_i \cdot \exp\{[E_I(x)-E_F]/kT\} \quad (525.1)$$

$$= P(-x_p)\cdot\exp\{-q[V_I(-x_p)-V_I(x)]/kT\} \quad (525.1A)$$

$$\simeq N_{AA}\cdot\exp\{-q[V_I(-x_p)-V_I(x)]/kT\} \quad (525.1B)$$

and

$$N(x) = n_i\cdot\exp\{[E_F-E_I(x)]/kT\} \quad (525.2)$$

$$= N(+x_n)\cdot\exp\{-q[V_I(+x_n)-V_I(x)]/kT\} \quad (525.2A)$$

$$\simeq N_{DD}\cdot\exp\{-q[V_I(+x_n)-V_I(x)]/kT\} \quad (525.2B)$$

using the dependence of $E_I(x)$ with x sketched in Fig.523.1(c) which gives the potential variation, $V_I(x) = E_I(x)/(-q)$, with x.

The hole concentration, P(x) in Fig.524.1(d), shows that holes are spilled over or transferred to the n-side of the p/n metallurgical boundary from the p-side as a result of hole diffusion since hole concentration is much higher on the p-side than on the n-side. Similarly, it also shows a slight spillover of the electrons into the p-side of the p/n metallurgical boundary from the n-side. We distinguish the term 'spilled over' from the term 'injection'. The term injection is reserved for the moving electrons or holes through a junction boundary by the action of an applied force. The term spillover will be used to denote the transfer of electrons and holes due to a concentration gradient causing diffusion at thermal and electronic equilibrium in the absence of an applied force.

Note that at $x=x_i$, $N(x_i)=P(x_i)=n_i$. This is also the position where the Fermi level, E_F, coincides with the intrinsic Fermi level, $E_I(x_i)$, as indicated in Fig.524.1(c). But x_i is not at the p/n boundary $x=0$.

The spillover of the electrons and holes into the space-charge layer lowers the peak electric field. The peak electric field occurs at $x=0$ in the depletion approximation as shown by (523.12). The true and the depletion-approximation electric fields are shown in Fig.524.1(e), labeled respectively by t and d. The peak magnitude of the true electric field is smaller than that of the depletion-approximation electric field due to the presence of electrons and holes in the space-charge layer which are neglected in the depletion approximation.

Furthermore, the presence of the mobile charges, N(x) and P(x), in the space-charge layer makes the boundaries, $x=-x_p$ and $x=+x_n$ ill-defined. This is demonstrated both by the mobile charge distributions in Fig.524.1(d) and the electric field variations in Fig.524.1(e) beyond these two boundaries. These exact or true results give a hint on how to make the first order analytical correction to the zeroth order depletion solutions which were just obtained.

530 DC ELECTRICAL CHARACTERISTICS OF A P/N JUNCTION

When a dc voltage is applied to the two terminals of a p/n junction diode, the balance of the drift and the diffusion currents at equilibrium is upset. A net current will pass through the p/n junction. A typical d.c. current-voltage characteristic of a p/n junction is shown in Fig.530.1(c) where V is the voltage applied to the p-type side of the the p/n junction relative to the n-type side and I is the current flowing into the terminal on the p-type side. Using this voltage polarity and current direction convention, the figure shows that when V is positive, there is a large positive current flowing from the p-side to the n-side. This is known as the **forward direction of current** and it is the direction of easy current flow. It is also known as the **forward bias** direction, i.e. a positive bias voltage is applied to the p-side relative to the n-side.

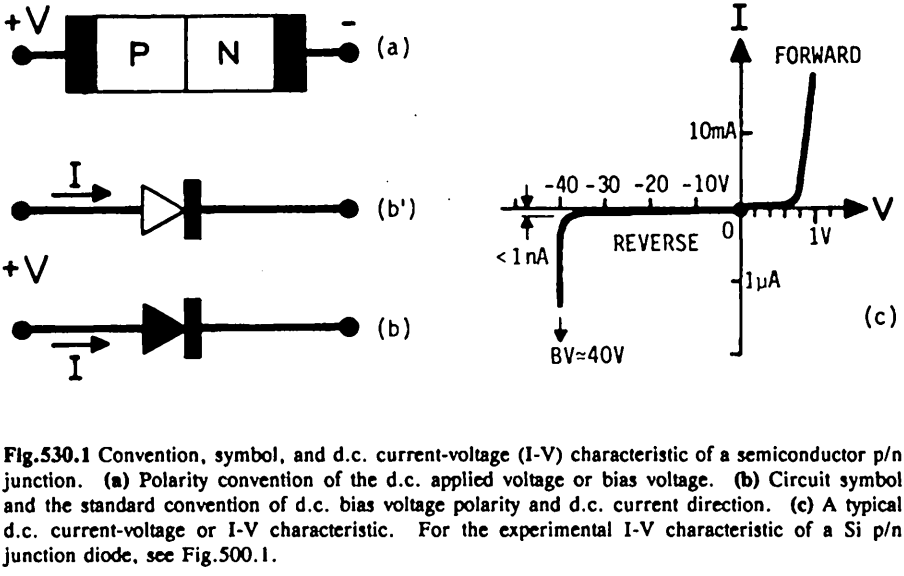


Fig.530.1 Convention, symbol, and d.c. current-voltage (I-V) characteristic of a semiconductor p/n junction. **(a)** Polarity convention of the d.c. applied voltage or bias voltage. **(b)** Circuit symbol and the standard convention of d.c. bias voltage polarity and d.c. current direction. **(c)** A typical d.c. current-voltage or I-V characteristic. For the experimental I-V characteristic of a Si p/n junction diode, see Fig.500.1.

When a negative d.c. voltage is applied to the p-side relative to the n-side or V is negative, Fig.530.1 shows a very small negative current ($I<0$) flowing through the p/n junction. This is known as the **reverse direction** of current. It is also known as the blocking direction, or the backward direction if we follow the antonyms, forward and backward. However, the terms blocking and backward are not used since their first letter, b, conflicts with that of bulk, back, body, base, and names of other parts of some transistor structures when the first letter of the names is used as the subscript to label current and voltage. The applied voltage is also known as the **reverse bias** when a negative d.c. bias voltage is applied to the p-side relative to the n-side.

The voltage scale for the forward or positive bias direction is greatly compressed in Fig.530.1(c) and in Fig.500.1(a). The forward current begins to increase rapidly when the forward voltage is greater than about 0.5 to 0.7 Volt in a typical silicon p^+/n or n^+/p junction diode. This rapid increase can be understood by **simple physical reasoning**. We recall that there is a built-in diffusion potential barrier of about $V_{bi}=V_D=0.7$ Volt across the p/n junction at equilibrium (i.e zero bias). This barrier prevents electrons and holes from crossing the p/n boundary easily since they have to overcome this potential barrier. If the forward bias is nearly equal to or greater than this built-in diffusion potential barrier, then the potential barrier across the p/n boundary layer is diminished or nearly vanishes. Thus, there is little or no potential barrier to prevent electrons and holes from crossing the p/n boundary, resulting in a large current flowing through the p/n junction.

Note that for the reverse or negative bias voltage direction shown in Fig.530.1(c), the voltage scale is greatly expanded compared with the scale of the forward current or positive voltage direction. Another view was given in Fig.500.1(b), especially the 10x curve. These reverse bias I-V plots show that even if the reverse voltage is several ten's of volts, we still have only a very low current flowing across the junction. This again can be understood by **simple physical reasoning** based on a consideration of the potential barrier across the p/n junction. When a negative voltage is applied to the p-side, the potential barrier is increased from V_{bi} to $V_{bi}+V_R$ where V_R is the magnitude of the reverse bias voltage, $V_R = -V = |V|$. Thus, very few electrons and holes will cross the p/n boundary because of the large potential barrier height. Only those few electrons and holes which are thermally generated near or inside the p/n junction space-charge layer can cross the junction and produce current. The thermal generation rate is very small and hence the reverse current is also very small. In a silicon p/n junction at room temperature (22°C), the maximum thermal generation or emission rate of electrons and holes at one impurity or defect generation center located inside the space-charge layer is about 1000 electrons and holes per second. We will show later how this estimate is obtained. A high power rectifier diode may have 10^{13}cm^{-3} generation centers in a space-charge layer of 10^{-4}cm thick and a junction area of $1\text{cm}\times1\text{cm}=1\text{cm}^2$ or a space-charge volume of $10^{-4}\text{cm}\cdot 1\text{cm}^2=10^{-4}\text{cm}^3$. This gives a total of $10^{13}\cdot 10^{-4} = 10^9$ generation centers and a total generation rate of $1000\cdot 10^9=10^{12}$ electron-hole/sec. The reverse current is given by the product (charge per electron) $\cdot$ (generation rate at one center) $\cdot$ (number of centers) $= 1.6\times10^{-19}\cdot 10^3\cdot 10^9 = 1.6\times10^{-6}\text{A} = 1.6\mu\text{A}$. For a p/n junction in a submicron VLSI circuit, the junction area is about $1\mu\text{m}\cdot 1\mu\text{m}=10^{-8}\text{cm}^2$ and the reverse current is reduced 10^8 times to $1.6\times10^{-6}\cdot 10^{-8}=1.6^{-14}\text{A}=16\text{fA}$ where f=femto=10^{-15}.

As the reverse bias voltage applied to the p/n junction is further increased in magnitude or becomes more negative, a voltage value will be reached, labeled BV in Fig.530.1(c), at which the reverse current will increase to infinity unless an external resistance or an internal series resistance is present to limit the current.

This is known as the **junction current breakdown condition** or simply **junction breakdown** and sometimes known as **current runaway**. The voltage BV is known as the **breakdown voltage**. The responsible physical phenomena is the interband impact generation of electron-hole pairs by energetic electrons and holes discussed in sections 3613 and 384. The term energetic is used to describe high kinetic energy of electrons and holes which is several times the thermal energy of kT≈25 meV at 300K. A significant number of energetic electrons and holes appear in the space-charge layer when a large reverse voltage is applied to the p/n junction, because the thermally generated electrons and holes will be accelerated by the high electric field to gain much kinetic energy before losing some of the acquired energy or velocity by phonon scattering, i.e. collision with the vibrating atomic cores. Thus, there will be many electrons and holes with kinetic energies several times the value of the energy gap, 1.2 eV in Si. This high kinetic energy favors the interband impact generation process of electron-hole pairs since these energetic electrons and holes will have more than the threshold energy needed to break a covalent bond and create an electron-hole pair. The generated electron-hole pairs will also be accelerated by the high electric field and they will generate additional pairs of electrons and holes. This **regenerative** process will diverge and give infinite electron-hole pairs and infinite current if one thermally generated electron (or hole) will generate one electron-hole pair while passing through the junction space-charge layer. This gives us the current breakdown or junction breakdown indicated in Fig.530.1(c). Thus, junction breakdown is also known as **avalanche breakdown** since the regenerative or positive feedback has created an avalanche of electron-hole pairs. This phenomenon is one of the **hot electron** (hot hole or **hot carrier**) effects. It is 'hot' because the effective temperature of an electron with several eV of kinetic energy, gained from having been accelerated through the potential drop of the space-charge layer, is very high. For example, the effective temperature of an electron with 1 eV kinetic energy is 1eV•(300K/25mV)=40•300K=12,000K while an average thermal electron has a kinetic energy of only 3kT/2 = (3/2)•25meV = 38meV at room temperature, T=300K≈27C.

The breakdown voltage, BV, can be varied from about 10 Volts to many hundreds of volts by controlling the dopant impurity concentrations in the p/n junction. Lower dopant impurity concentration gives thicker space-charge layer, as indicated by (523.19), where V_{bi} is now replaced by $V_{bi}+V_R$. It also gives lower electric field as indicated by (523.21) and (523.22) where V_{bi} is also replaced by $V_{bi}+V_R$. Hence, at a lower dopant impurity concentration, it would take a higher reverse voltage to reach the breakdown electric field condition to give infinite current or current breakdown.

In the early days of semiconductor technology, the dopant impurity of p/n junction was not uniformly distributed and localized physical defects were frequently present. These imperfections gave localized high electric fields, causing localized current divergent spots across the junction area. If a high breakdown current is passed through the diode, light is generated at these current-concentration spots since some of the energetic electrons and holes may recombine and give off light. Due to the high current, some spots may also be heated to high temperatures giving off a spotty glow. The scene resembles the aerial view of the city lights at night. Impurity clusters and physical defects still occur today when production processes go astray or out of control. This is one source of low production yield, reduced operating life, higher-cost, and lower-profit.

As the impurity doping concentration increases, the junction space-charge layer becomes thinner and the peak electric field increases so that the current diverges at a lower reverse bias. When the breakdown voltage is lowered to about 6 or 7 Volts by increasing the dopant impurity concentration, another physical phenomenon becomes responsible for the rapid increase of the reverse current at larger reverse voltages. The interband impact generation phenomenon just described becomes ineffective since the space-charge layer is now too thin to give the electron or hole the chance to generate an electron-hole pair by impact while passing through the space-charge layer. The new phenomenon that causes the rapidly increasing reverse current with increasing reverse voltage is **quantum mechanical tunneling**. The p/n junction potential barrier is now so thin that a electron on the n-side will have a high probability to tunnel through the barrier and end up in a hole on the p-side of the barrier (or a hole will tunnel across the barrier and end up in an electron). This is described in section 36n0 and illustrated by the transitions (a) and (b) shown in the tunneling transition energy band diagram Fig.36n0. Higher doping impurity density gives thinner potential barrier and higher tunneling probability, and hence larger tunnel current. However, tunneling is **not a regenerative process** in contrast to the interband impact generation of electron-hole pairs which is a regenerative process. Thus, the tunnel current will rise rapidly as the reverse applied voltage is increased beyond a certain value, but the current will not runaway or diverge at some bias voltage like the interband impact generation phenomenon. A true electronic breakdown of current will not occur due to tunneling.

In the early 1950's, this tunneling current was thought to be responsible for the diode breakdown current. The term, **Zener breakdown**, was used by McAfee, Ryder, Shockley and Sparks in 1951 after Clarence Zener who proposed in 1934 that interband tunneling may account for the dielectric breakdown phenomenon observed in materials. The term, Zener diode, was universally applied to all semiconductor diodes designed to operate at the breakdown voltage as voltage regulators, including even the higher breakdown voltage diodes whose current breakdown is caused by interband impact collision and not quantum mechanical tunneling. Shockley told me during my apprentice with him that this was one of the two theories he erred at Bell Labs. (The other was the non-existent negative resistance he recorded in his notebook pages on the theory of the junction-gate field-effect he invented.) But Zener tunneling current is indeed an important component of reverse diode current albeit it is incapable of causing an electronic avalanche current breakdown or current runaway.

Since the current at breakdown can become very high (tunneling) or infinite (interband impact generation of e-h pair) and can be highly localized at defect sites where the electric field is high, the silicon at these sites can be heated to very high temperatures. At higher temperatures, the diode current would be higher due to higher thermal generation rate of electrons and holes. This is a regenerative electronic-thermal cycle which would increase the temperature to the melting point of silicon and cause the physical destruction of the p/n junction.

531 Energy Band Diagram of a Biased p/n Junction

The procedures to construct the energy band diagrams under reverse and forward bias voltages are described in the following paragraphs.

Reverse Bias

We shall first obtain the thickness of the space-charge layer as a function of the bias voltage. It seems intuitively obvious that for reverse bias, the depletion approximation should apply very well. In fact, it should be better than at zero bias which we used to derive the equilibrium properties of p/n junctions in section 523. The reason is that electrons are drawn to the n-side and holes to the p-side by the reverse bias, consequently the electrons and holes are depleted in the space-charge layer. Thus, to obtain the space-charge layer thickness at a reverse bias voltage, we only need to modify the zero-bias expression of the potential barrier height across the p/n junction given by (523.19). At zero bias, the barrier height is V_{bi}. When a reverse bias of magnitude V_R is applied, the barrier height increases to $V_{bi}+V_R$. Using the polarity convention of positive for the p-type terminal shown in Fig.530.1(a) and (b), then $V = -V_R$. The thickness of the space-charge layer can be obtained from (523.19) if we replace V_{bi} by $V_{bi}+V_R$:

$$x_{pn} = \sqrt{2\varepsilon_s(V_{bi}+V_R)/qN_M} \qquad (531.1)$$

where $N_M=N_{AA}N_{DD}/(N_{AA}+N_{DD})$. The maximum electric field from evaluating (523.21) at $x=0$ is

$$\mathcal{E}_{max} = \sqrt{2qN_M(V_{bi}+V_R)/\varepsilon_s} \qquad (531.2)$$

$$= 2(V_{bi} + V_R)/x_{pn}. \qquad (531.2A)$$

The replacement of V_{bi} by $V_{bi}+V_R$ can be visualized by examining the energy band diagram under a reverse bias which is shown in Fig.531.1(a). The applied voltage causes a current to flow so that we can no longer use a Fermi level since it is defined only at thermal equilibrium. However, we can introduce a quasi-Fermi level whose gradient gives the current. This definition was first introduced in section 331. From (331.1) and (331.1A), the electron current is

$$J_{Nx} = + \mu_n N(dF_N/dx) = - q\mu_n N(dV_N/dx). \qquad (531.3)$$

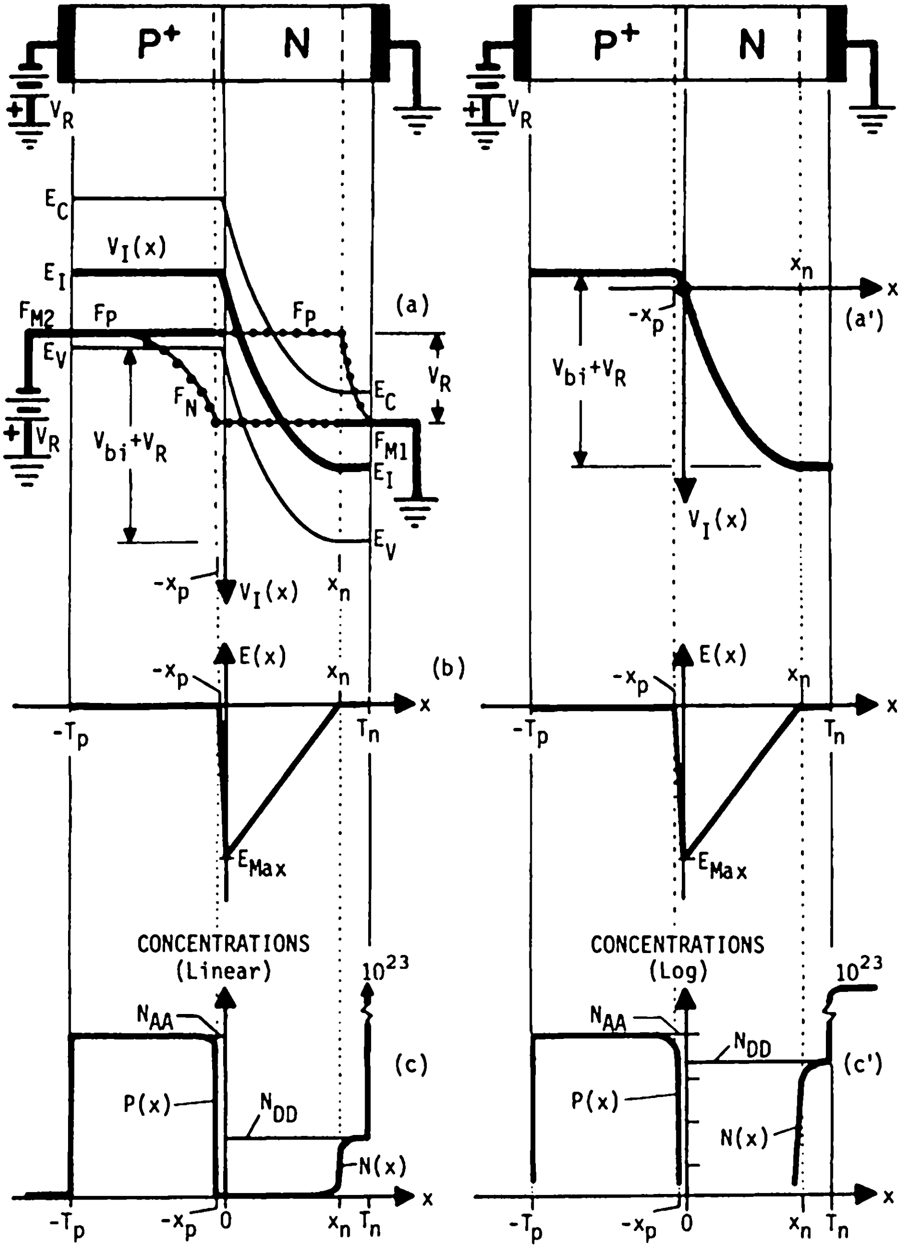

P+
N
V_R
E_C
$V_I(x)$
E_I
F_{M2}
F_P
E_V
F_N
$V_{bi}+V_R$
(a)
V_R
E_C
F_{M1}
E_I
E_V
$-x_p$
$V_I(x)$
x_n
(a')
x
E(x)
(b)
$-T_p$
T_n
E_{Max}
CONCENTRATIONS (Linear)
CONCENTRATIONS (Log)
10^{23}
N_{AA}
N_{DD}
P(x)
N(x)
(c)
(c')
0

Fig.531.1 P/N junction diode under reverse bias. **(a)** Energy band diagram and **(a')** electric potential. **(b)** Electric field. **(c)** Concentration of electrons, holes, acceptors and donors in linear and **(c')** expanded semi-logarithmic scales.

The corresponding electron concentration under nonequilibrium was also similarly defined by the quasi-Fermi potential for electrons given by (331.5)-(331.7)

$$N(x) = N_C \exp[-(E_C-F_N)/kT] \tag{531.4A}$$

$$= n_i \exp[(F_N-E_I)/kT] \tag{531.4B}$$

$$= n_i \exp[q(V_I-V_N)/kT]. \tag{531.4C}$$

Similarly we had (331.3) and (331.3A) for hole currents, and (331.8)-(331.10) for hole concentrations. These are,

$$J_{Px} = - q\mu_p P(dV_P/dx) \tag{531.5}$$

$$P(x) = N_V \exp[-(F_P-E_V)/kT] \tag{531.6A}$$

$$= n_i \exp[(E_I-F_P)/kT] \tag{531.6B}$$

$$= n_i \exp[(V_P-V_I)/kT]. \tag{531.6C}$$

Note in Fig.531.1(a) that the total variation of the electron quasi-Fermi potential, $V_N(-T_p)-V_N(T_n)$, hole quasi-Fermi potential, $V_P(-T_p)-V_P(T_n)$, and the separation of the electron and hole quasi-Fermi potentials across the p/n junction transition layer, $V_P(-x_p<x<x_n)-V_N(-x_p<x<x_n)$, are all equal to the voltage applied, $-V_R$. This is a logical result since in the n-region far away from the p/n junction boundary, the electron (majority carrier) quasi-Fermi level (i) should be nearly equal to its equilibrium value, and (ii) is nearly constant and joins onto the electron quasi-Fermi level of the electrons in the battery which supplies the reverse applied voltage. Similarly, the hole quasi-Fermi level in the p-region far away from the p/n boundary is nearly (i) equal to its equilibrium value and (ii) coincident with the electron quasi-Fermi level in the metal which makes a low-resistance ohmic contact to the p-region at $x=-T_p$. The electron quasi-Fermi level in the metal coincides with electron quasi-Fermi level of the battery's negative electrode which is at a potential of V_R lower than the battery's positive electrode. This gives the self-consistent description of the potential and quasi-Fermi potential variation through the entire loop containing the p/n junction diode, the metal connection wires and the battery which supplies the applied voltage.

The spatial constancy of the majority carrier quasi-Fermi potential or level, such as $V_P(x)$ in the p-region and $V_N(x)$ in the n-region, can also be proved qualitatively by the current density equations written as a gradient of the quasi-Fermi potentials, (531.3) and (531.5). The spatial constancy is a result of the high concentration of the majority carriers, since from these equations we have $dV_N/dx = J_{Nx}/(-q\mu_n N)$ and $dV_P/dx = J_{Px}/(-q\mu_p P)$ so $dV_N/dx \to 0$ when N is very large and the electron current J_{Nx} is not too large, and similarly for dV_P/dx.

Forward Bias

When a positive or forward voltage is applied to the p-side of the p/n junction relative to the n-side, holes are **injected** from the p-side into the n-side because of the potential energy of holes on the p-side is increased by the applied positive voltage while it is decreased on the n-side. This is illustrated in Fig.531.2(a). The quasi-Fermi levels, F_P and F_N are split by the amount -qV. The corresponding split of the quasi-Fermi potentials is $V_P - V_N = V$.

The shift of the potential energy causes the holes to flow or to be **injected** from the p-side to the n-side. A similar reasoning applies to electrons which are injected from the n-side into the p-side. This is known as **minority carrier injection** since the holes injected from the p-side into the n-side are the minority carriers in the n-side. Similarly electrons from the n-side injected into the p-side are the minority carriers in the p-side. These minority carrier injections cause a forward or positive current to flow from the p-side to the n-side of the p/n junction.

The injection of the carriers to the other side is illustrated by the spread of their densities, P(x) and N(x), shown in Figs.531.2(c) and (c'). An expanded view is shown in Fig.531.2(d) on the following page where the electron concentrations at the four boundaries, $x=-T_p$, $-x_p$, $+x_n$ and $+T_n$, are also given. These are the boundary conditions which will be used to derive the analytical solutions of the currents in the next two sections. The magnitude of the current depends on the **rate of recombination** of the injected minority carriers with the majority carriers. The recombination rate affects and controls the rate of minority carrier injection because injection itself cannot maintain a steady-state current, since the injected carriers must find a sink in order to become trapped and the sink either cannot be full or must have a channel to leak off the captured electrons. The leak comes from capturing holes by the trap since the captured holes will then recombine with the captured electrons. The recombination energy is carried away by the thermal vibration of the silicon atoms, i.e. phonons.

The space-charge or p/n junction transition layer, extended from $-x_p$ to $+x_n$, is even less well defined than at V=0. This is indicated by the spreading out of the holes and electrons into the transition layer shown in Figs.531.2(c), (c') and (d), especially on the linear scale in Fig.531.2(c). Thus, the thickness computed from the depletion model, (523.19), is no longer a good approximation. However, an accurate analytical formulae is not available. Thus, the depletion formulae for the space-charge layer thickness, (523.19), and the maximum electric field, (523.21), are frequently used. Replacing V_{bi} by V_{bi}-V where V is the forward applied d.c. voltage, these are

$$x_{pn} = \sqrt{2\varepsilon_s(V_{bi}-V)/qN_M} \tag{531.7}$$

and

$$E_{max} = \sqrt{2qN_M(V_{bi}-V)/\varepsilon_s}. \tag{531.8}$$

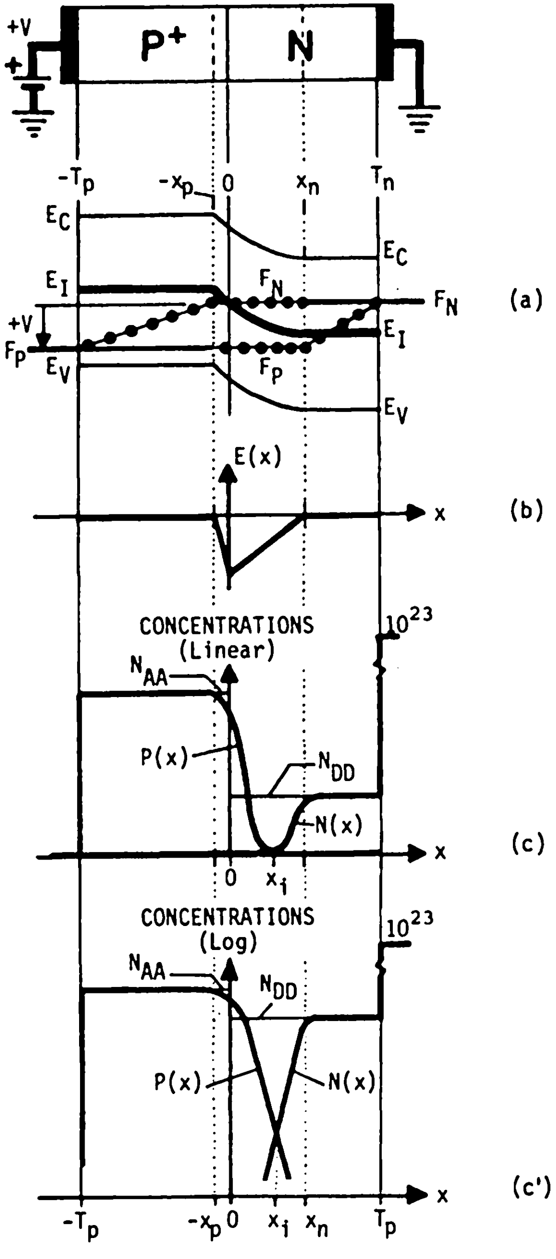

+V
P+
N
-Tp
-xp
0
xn
Tn
EC
EI
FN
+V
FP
EV
FP
(a)
E(x)
x
(b)
CONCENTRATIONS
(Linear)
10²³
NAA
P(x)
NDD
N(x)
x
(c)
0
xi
CONCENTRATIONS
(Log)
10²³
NAA
NDD
P(x)
N(x)
x
(c')
-Tp
-xp
0
xi
xn
Tp

Fig.531.2 Forward biased p/n junction. **(a)** E-x energy band diagram and electric potential (dark-heavy curve). **(b)** Electric field. **(c)** $P(x)$, $N(x)$, $N_{AA}(x)$ and $N_{DD}(x)$ in linear scale and **(c')** in expanded semilogarithmic scale. (d) See further expansion of (c') on next page.

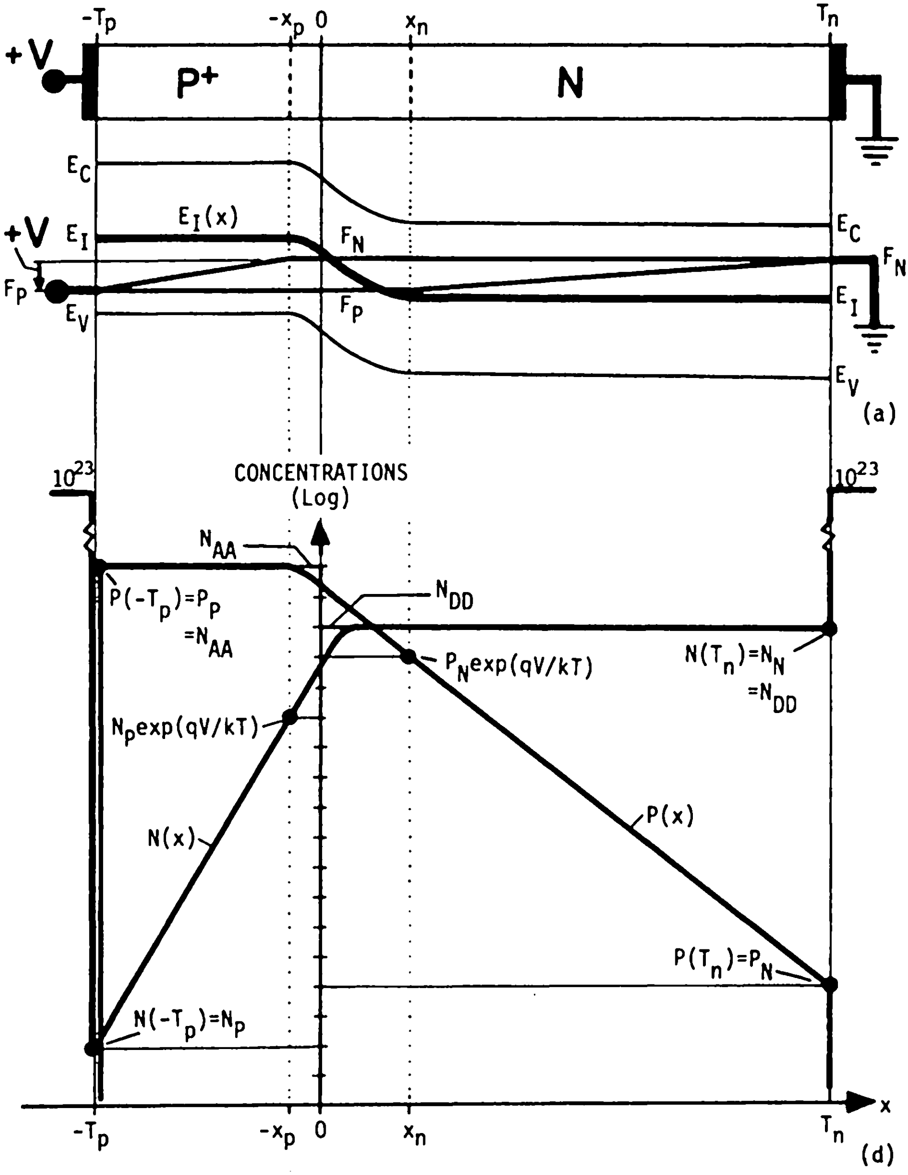

-Tp
-xp
0
xn
Tn
+V
P+
N
EC
EI(x)
+V
EI
FP
EV
FN
FP
EC
FN
EI
EV
(a)
10^23
CONCENTRATIONS
(Log)
10^23
NAA
P(-Tp)=Pp
=NAA
NDD
PNexp(qV/kT)
N(Tn)=NN
=NDD
Npexp(qV/kT)
N(x)
P(x)
P(Tn)=PN
N(-Tp)=Np
x
-Tp
-xp
0
xn
Tn
(d)

Fig.531.2 Forward biased p/n junction. **(a)** E-x energy band diagram and electric potential (dark-heavy curve). (b)-(c') on previous page. **(d)** Expanded (c') in semilog of P(x) and N(x) with boundary value equations labeled.

532 The Shockley Diode Equation

The energy band diagram shown in Fig.531.2 is used to analyze the minority carrier injection current. We shall work out this problem using several simplifying assumptions. Consider the n-type side. We shall assume that the injected hole concentration is very low compared with the equilibrium concentration of electrons on the n-side so that the charge neutrality condition still nearly holds. We shall also assume that the drift current of the holes (minority carriers) is small compared with the diffusion current. This is a good approximation since the minority carrier density is very small and the electric field in this neutral region is also very small so that the minority carrier drift current, $q\mu_p PE$, would also be very small.

Thus, the steady-state hole continuity equation and the hole current equation given by the Shockley equations (340.2) and (340.4) have the following simplified one-dimensional forms.

$$q\partial P/\partial t = 0 = -\,dJ_P/dx + q(G_P - R_P) \tag{532.1}$$

$$J_P = q\mu_p PE - qD_p dP/dx \approx -\,qD_p dP/dx. \tag{532.2}$$

We also assume that the net rate of generation over recombination, G_P-R_P, can be approximated by a constant hole lifetime or by a linear recombination law. It was demonstrated in section 372 that a constant lifetime can be defined. For the Shockley-Read-Hall thermal recombination-generation process at a defect or impurity center, the constant lifetime given by (372.4) was $\tau_p = 1/c_{pe}{}^{t}N_{TT} \equiv \tau_{p0}$. Using this linear recombination law, we have

$$G_P - R_P = (P_E - P)/\tau_p$$

which can be substituted into (532.1) to give

$$dJ_P/dx = q(G_P - R_P) = q(P_N - P)/\tau_p \tag{532.3}$$

where $P_N = P_E$ is the equilibrium hole concentration or minority carrier concentration on the n-side. This can be combined with (532.2) to give

$$D_p d^2P/dx^2 = (R_P - G_P) = (P - P_N)/\tau_p \tag{532.4}$$

or

$$d^2P/dx^2 = (P - P_N)/(D_p\tau_p) \equiv (P - P_N)/L_p{}^2. \tag{532.5}$$

This is a second order linear differential equation with constant coefficient if both D_p and τ_p are independent of position. In the second form, (532.5), we define a **diffusion length for minority carriers** (holes) by

$$L_p = \sqrt{D_p\tau_p}. \tag{532.6}$$

The general solution of (532.5) is

$$P(x) - P_N = A\exp(x/L_p) + B\exp(-x/L_p) \qquad (532.7)$$

and the hole current or minority carrier current at a position x is

$$\begin{aligned} J_P(x) &= -qD_p dP/dx \\ &= -qD_p[(A/L_p)\exp(x/L_p) - (B/L_p)\exp(-x/L_p)] \\ &= -q(D_p/L)[A\exp(x/L_p) - B\exp(-x/L_p)]. \end{aligned} \qquad (532.8)$$

The two constants, A and B, must be evaluated using two boundary conditions for a given junction geometry, dopant impurity profile and applied voltage. As illustrated in Fig.531.2(a), the diode can be **partitioned into three layers**: the quasi-neutral p+ layer on the left, the space-charge layer in the middle, and the quasi-neutral n layer on the right. Thus, the constants A and B can be evaluated using the two boundary conditions at the two interfaces of each of these three layers. For example, the two boundary conditions of the quasi-neutral n layer on the right are at the planes $x=x_n$ and $x=T_n$.

To illustrate, we take the simplest case of a long or thick p+/n diode on a thick silicon or semiconductor wafer. The wafer is sufficiently thick so that its n-layer thickness, T_n, is many times the minority carrier (hole) diffusion length, L_p, i.e. $T_n >> L_p$. Consider the **first boundary condition** which is at the n-type surface of the wafer, $x=+T_p$. The hole (minority) concentration is equal to its equilibrium value, $p(x=+T_n)=P_N$ as indicated in Fig.531.2(d), since the influence of the voltage applied to the p/n junction space-charge layer boundary would not be felt so far away at $x=T_n$. Then, $A=0$ in (532.7), and (532.7) and (532.8) become

$$P(x) = P_N + B\exp(-x/L_p) \qquad (532.7n)$$

and

$$J_P(x) = -qD_p dP/dx = (-qD_p)[-(B/L_p)]\exp(-x/L_p). \qquad (532.8n)$$

Consider the **second boundary condition** which is at the n-side surface of the space-charge layer, $x=+x_n$. The hole concentration is raised by the Boltzmann factor, $\exp(qV/kT)$, above its equilibrium value, P_N, because the applied voltage, V, appears across the space-charge layer and raises the local potential at x_n by V volt. This is illustrated in the carrier density diagram shown in Figs.531.2(c), (c') and (d). This boundary condition for the hole concentration can be readily derived from the nonequilibrium relationship for the hole concentration, $P(x) = n_i\exp\{q[V_P(x)-V_I(x)/kT]\}$. Evaluating P(x) at $x=x_n$ using the forward-bias energy band diagram shown in Fig.531.2(a), the second boundary condition is then

$$P(x=x_n) = P_N\exp(qV/kT) \qquad (532.9)$$

which can also be deducted from and is labeled in Fig.531.2(d). Using this boundary condition in (532.7) and the first boundary condition $P(T_n)=P_N$ which gave $A=0$, the second constant is then

$$B = P_N[\exp(qV/kT) - 1]\exp(x_n/L_p).$$

The spatial variations of the concentration and current density of holes with position in the n-type quasi-neutral layer are then given by

$$P(x) = P_N + P_N[\exp(qV/kT) - 1]\exp[-(x-x_n)/L_p] \qquad (532.10)$$

and

$$J_P(x) = (qD_pP_N/L_p)[\exp(qV/kT) - 1]\exp[-(x-x_n)/L_p]. \qquad (532.11)$$

The hole current flowing into the n-type quasi-neutral region from the space-charge layer, $J_P(x_n)$, is then given by

$$J_P(x_n) = (qD_PP_N/L_p)[\exp(qV/kT) - 1]. \qquad (532.12)$$

This is **half** of the Shockley Diode Equation. The other half comes from making a similar analysis for the minority carriers, electrons, in the quasi-neutral p+ layer on the left side of the p/n junction illustrated in Fig.531.2(d). The complete Shockley diode equation is then given by

$$\begin{aligned} J &= J_P(+x_n) + J_N(-x_p) \\ &= [(qD_pP_N/L_p) + (qD_nN_P/L_n)][\exp(qV/kT) - 1] \qquad (532.13) \\ &= J_1[\exp(qV/kT) - 1] \quad \textbf{Shockley Diode Equation} \qquad (532.14) \end{aligned}$$

The coefficient J_1 contains the two current components: one each from the p-type and n-type quasi-neutral layers. It is defined by

$$J_1 = J_{P0} + J_{n0} = (qD_pP_N/L_p) + (qD_nN_P/L_n). \qquad (532.15)$$

J_1 is known as the **reverse saturation current** or **junction saturation current**. At large reverse biases or large negative V, exp(qV/kT) can be dropped compared with 1 in the term [exp(qV/kT) - 1] of (532.14), then the Shockley diode current saturates to a voltage-independent constant value, $-J_1$.

533 Physics of the Shockley Diode Equation

A detailed analysis of the Shockley diode equation will enhance our understanding of the physics of semiconductor diode tremendously. We consider the saturation current first. The saturation current of the Shockley diode can be changed to a different form using $L_p = \sqrt{D_p\tau_p}$ and $L_n = \sqrt{D_n\tau_n}$ in (532.15). The new forms are listed in the second positions of each of the following two equations.

$$J_{p0} = qD_pP_N/L_p = qL_pP_N/\tau_p \tag{533.1A}$$

and

$$J_{n0} = qD_nN_P/L_n = qL_nN_P/\tau_n. \tag{533.1B}$$

The two alternative forms of each of the above two saturation current components emphasize the **two mechanisms** which control their magnitude. The first form for holes in (533.1A) emphasizes its minority carrier (hole) **diffusion origin** in the n-type quasi-neutral layer. This is indicated by qD_pP_N/L_p = (charge) x (minority carrier diffusivity) x (minority carrier concentration) / (minority carrier diffusion length). The last two factors, (minority carrier concentration) / (minority carrier diffusion length), is in the unit of concentration gradient signifying diffusion.

The second form of the hole saturation current given in (533.1A) emphasizes its minority carrier (hole) **recombination origin** in the n-type quasi-neutral layer. This is indicated by qL_pP_N/τ_p = (charge) x (minority carrier diffusion length) x (minority carrier concentration) + (minority carrier lifetime). The last two factors, (minority carrier concentration) + (minority carrier lifetime), is in the unit of volume density of recombination rate of minority carriers. The minority carrier diffusion length, L_p, also enters this recombination interpretation through the recombination volume. Thus, the recombination volume is given by the product of a layer thickness of L_p and a junction area of A. To describe: the Shockley diode current is a result of the recombination of minority carriers with majority carriers in the quasi neutral n-type volume of L_pA at a rate of P_N/τ_p for holes, and in the quasi neutral p-type volume of L_nA at a rate of N_P/τ_n for electrons.

These two interpretations show that the Shockley diode equation is not just due to minority carrier diffusion. **Diffusion alone is insufficient to produce a steady-state current** since the carriers arriving at a low density region will increase its concentration, making the concentrations at the initially low and high concentration regions equal and the concentration gradient zero. The diffusion current then ceases after this initial transient. Thus, the often used diffusion interpretation is incomplete. In order to have a steady-state current, one must have a **sink** for the arriving minority carriers. In the case of the long or thick diode used in the Shockley model, the sink is the recombination-generation-trapping centers **distributed** over the **entire** quasi-neutral n-type and p-type layers. For a short or thin diode which has no bulk recombination center, the sink are the two metal-semiconductor contacts where the recombination rate is infinite or very large. So, the correct terminology for the Shockley diode current is **diffusion-recombination** current. Recombination is paramount for having a steady-state current. In this example, it is the Shockley-Read-Hall **thermal recombination mechanism** at the impurity and defect centers distributed in the n-type and p-type quasi-neutral layers that maintains the steady-state d.c (forward) current.

In the above interpretation and analyses of the physical origin of the Shockley diode current, we were not concerned with the sign. The sign is significant since it brings in a third physical mechanism which we have learned in chapter 3, the **thermal generation of electrons and holes**. This can be demonstrated from the Shockley diode equation as follows. For large reverse bias, $V<<0$, we have

$$J(V<<0) \approx - J_1 = - J_{p0} - J_{n0}. \qquad (533.2)$$

The negative sign in the diffusion interpretation indicates that the current is flowing in the negative x direction since the concentration of the minority carriers is reduced to zero at the two boundaries of the space-charge layer, $x=-x_p$ and $x=+x_p$. It is zero because the large reverse bias voltage and the high electric field in the space-charge layer draw the minority carrier out of the space-charge layer. This gives a concentration gradient for holes in the negative direction and for electrons in the positive direction, resulting in the negative sign. Thus, the physics of the diffusion interpretation is still valid.

The negative sign in the recombination interpretation introduces a third origin, the thermal generation mechanism. For example, in $-J_{p0} = -qL_p(P_N/\tau_p) = qL_p(-P_N/\tau_p)$, the negative sign is grouped with $(-P_N/\tau_p)$ or $-\tau_p$ since a negative carrier density $-P_N$ is physically meaningless. The negative recombination lifetime, $-\tau_p$, means that it is the lifetime or reciprocal rate of the inverse process, the carrier **generation** process. This is frequently termed the **generation lifetime** when it is measured under a reverse bias experiment. In our simple model, the recombination and generation lifetimes are assumed equal. In practice, this may not be the case because the electron (and hole) capture or recombination rate may not equal the electron emission (and hole) or generation rate at a generation-recombination-trapping center. The difference can come from several factors. For the case of only one species of centers in the quasi-neutral layers, the capture and emission rates would be the same. But for multi-species of centers in the quasi-neutral layers, one species may be more effective for recombination under the forward bias condition while a second species may be more effective for generation under the reverse-bias condition. The details of these difference can be studied based on the foregoing physics and the basic theories given in chapter 3. They are left as a subject for a second course on solid state devices.

The Shockley diode equation can be dissected to give additional physical insights and physics. We proceed in three ranges of applied voltage, the choice of which can be readily made by looking at the Shockley diode equation. Focusing on the exponential voltage dependent term in (532.14), we immediately recognize that at large reverse voltages, $\exp(V/kT)=\exp(q|V|/kT)$ would be very small compared with 1 and can be deleted. For large forward voltage, $\exp(qV/kT)$ would be very much larger than 1 and would dominate. The quantitative conditions of these two limits can be established by requiring that the exponential term is 100 times larger or smaller than 1 which translates to an exponent value of 4.6. A value of 4.0 is used since 4kT/q≈4x25mV=100mV at room temperatures (T≈290K or about 17C). Thus, (qV/kT) > 4 or < -4, and in these two ranges, we have the following approximation solutions for the diode current

$$J = +\,J_1\exp(qV/kT) \qquad V > 4(kT/q) = 0.1\mathrm{V}, \qquad (533.3)$$

and

$$J \approx -\,J_1 \qquad V < -4(kT/q) = -0.1\mathrm{V}. \qquad (533.4)$$

The forward current, (533.3), arises from diffusion and recombination of the minority carriers **injected** into the two quasi-neutral layers which we have just discussed. This is also known as the **injection current**. The reverse current, (533.4), arises from generation of minority carriers in the two quasi-neutral layers within one diffusion length of the space-charge layer which we have also just discussed. This is also known as the **leakage** or **saturation current**. The asymptotic solutions given above show that the diode is a **highly nonlinear circuit** element only when the voltage applied extends over a range greater than ±100 mV. The nonlinearity diminishes when the voltage swing is small.

The near-zero small bias voltage range is the third range. It has a very important fundamental significance in basic semiconductor device physics and diode application. It is not as obvious as the two large voltage ranges just discussed. To delineate this low voltage range, we again focus on the exponential voltage dependent term, exp(qV/kT), and note that the power series expansion of an exponential function is given by

$$\exp(Z) = 1 + Z/1! + Z^2/2! + Z^3/3! + \ldots \qquad (533.5)$$

This shows that a linear approximation is achieved if the second and higher order terms in the series can be dropped compared with the linear term. Let the criterion be 10% error (instead of the 1% assumed for the two asymptotic limits just discussed), then

$$(10)\cdot Z^2/2! < Z/1! \qquad \text{or} \qquad |Z| < 2/10 \qquad (533.6)$$

which can be used in the exponential term, exp(qV/kT), to give $|qV/kT|<(1/5)$ or $|V| < kT/5q = 25\mathrm{mV}/5 = 5$ mV. The linear range is then $-5\mathrm{mV} < V < 5\mathrm{mV}$ at about 17C or 290K. In this low voltage range, the Shockley diode current-voltage characteristic is linear or ohmic. It is given by

$$J = J_1(qV/kT) = (qJ_1/kT)V \qquad \text{for } -5\mathrm{mV} < V < 5\mathrm{mV}. \qquad (533.7)$$

This is an important result. In applications, it states that the nonlinear or rectification property of the Shockley diode ceases when the magnitude of the applied voltage is less than about 5 mV. In basic physics, it states that a solid state device or material will maintain a linear electrical or current-voltage characteristic when the perturbation or disturbance causes less than about kT/5q or 0.2(kT/q) potential fluctuation, if the current is due to thermal diffusion, drift, generation-recombination of electrons and holes. For diodes whose conduction is based on other mechanisms such as quantum mechanical tunneling, the linear property may persist only in a much smaller voltage range than the thermal voltage range of -kT/q to +kT/q and may even not exist at all in some cases.

534 Numerical Example of a Shockley Diode

Let us plug in some numbers to get a feel of the order of magnitude of the current one can expect from a Shockley diode. Suppose we have a silicon diode with $\mu_n = 1000 cm^2/V\text{-}s$ and $\mu_p = 360 cm^2/V\text{-}s$ at 297K, then $D_n = (kT/q)\mu_n = 25cm^2/s$ and $D_p = (kT/q)\mu_p = 9cm^2/s$. Let us assume that the concentrations are $N_{AA} = 10^{18}cm^{-3} \approx P_P$, $N_{DD} = 10^{14}cm^{-3} \approx N_N$, and $n_i = 10^{10}cm^{-3}$ at about 297K, then

$$P_N = 10^{20}/10^{14} = 10^6 \ cm^{-3},$$

and

$$N_P = 10^{20}/10^{18} = 10^2 \ cm^{-3}. \qquad (534.1)$$

Assume a 1-μs lifetime for both electrons and holes, then the minority carrier diffusion lengths are

$$L_n = \sqrt{D_n\tau_n} = \sqrt{25x10^{-6}} = 5x10^{-3} \ cm = 50 \ microns,$$

and

$$L_p = \sqrt{D_p\tau_p} = \sqrt{9x10^{-6}} = 3x10^{-3} \ cm = 30 \ microns. \qquad (534.2)$$

These can be substituted into (532.15) to give

$$J_{p0} = 1.6x10^{-19}x \ 9x10^6/3x10^{-3} = 4.8x10^{-10} \ A/cm^2, \qquad (534.3A)$$

$$J_{n0} = 1.6x10^{-19}x25x10^2/5x10^{-3} = 8.0x10^{-14} \ A/cm2, \qquad (534.3B)$$

and

$$J_1 = J_{p0} + J_{n0} = 4.8x10^{-10} + 8.0x10^{-14} \ A/cm2 \approx J_{p0}. \qquad (534.3C)$$

This shows that the diffusion current is mainly determined by minority carrier recombination-generation-diffusion in the lower doped n-quasi-neutral layer.

535 The Sah-Noyce-Shockley Diode Equation

What we have neglected in the Shockley Diode Equation, (532.14), is the current from the third or the space-charge layer of the p/n junction diode shown in Figs.531.2(a) $-x_p \le x \le x_n$. In deriving the Shockley diode equation we considered only the two quasi-neutral layers and not the space-charge layer. Electrons and holes are continually generated and recombine at the impurity and defect centers located in the space-charge layer and hence are expected to produce a current. At thermal equilibrium or zero applied voltage, the rate of generation must balance the rate of recombination. When a voltage is applied, this balance is destroyed and a net rate of recombination over generation will occur under forward bias and a net rate of generation over recombination will occur under reverse bias. The imbalance between recombination and generation in the space-charge layer gives rise to an additional diode current.

Diode current from the junction space-charge layer was neglected in Shockley's 1949 diode theory because at that time, only Ge crystal of high physical perfection and high purity could be grown. Diodes made from Ge crystal had I-V which followed faithfully the Shockley diode equation. With the advent of Si crystal growth technology in the early 1950's, Si p/n junction diodes could then be made. It was found that the Shockley diode equation could not explain Si diodes' I-V characteristics in an extended range of applied voltage, from large negative bias to small forward bias. In particular, the first observations of the discrepancies were: (1) the reverse current of the silicon p/n junction diode does not saturate at large reverse bias as predicted by - J_1 in (533.4) of the Shockley diode equation, and (2) the forward current does not follow the exponential law, exp(qV/kT), at low forward bias.

The observed reverse current was found to increase with reverse bias voltage following a power law, $(|V|)^m$, while the observed forward current was found to vary as exp(qV/nkT) where $n > 1$. The factor n (not electron concentration) has been known to circuit and device designers and engineers as the **diode ideality factor**. During the development of the space-charge layer recombination theory in 1956 (published in 1957) by Sah, Noyce and Shockley, it was recognized that current from electron-hole generation and recombination in the space-charge layer is more important in materials with large energy gap. As a consequence, the theory has been applied to and guided the development of diodes in other semiconductors, such as GaAs, InP, CdS, ZnS, and others. It has also been applied to the space-charge layers of other Si device structures, such as the metal/semiconductor, MOS diodes, and MOS transistors in Si integrated circuits, and recently, compound semiconductor heterojunction diodes. In silicon MOS integrated circuits, generation current from the space-charge layer at reverse bias is a principal component of standby current which must be taken into account to design the standby power dissipation and other properties such as the refresh rate of the dynamic random access memory (DRAM) cell. In Si bipolar transistors, electron-hole recombination in the space-charge layer of the forward biased emitter-base junction prevents minority carrier from being injected into the base layer which limits transistor current gain at low currents. Special transistor geometry and fabrication techniques had to be developed to produce super-beta bipolar junction transistor in order to give high performance and high input impedance bipolar Operational Amplifiers

(OP AMP). Several of these applications will be described more quantitatively in latter chapters on MOS and bipolar junction transistors.

The theory of the space-charge-layer recombination-generation current can be readily derived using the second interpretation we have just developed for the current coefficient, J_1, of the Shockley diode equation. There are three keys to be used in the derivation. (i) At zero applied voltage or thermal equilibrium, the electron and hole concentrations are equal at the intrinsic point, $x=x_i$, and are given by intrinsic carrier density n_i. (ii) The electron and hole concentrations at this position increase and decrease with applied voltage but remain about equal so that $N(x_i)=P(x_i) \propto n_i \exp(qV/2kT)$. The factor '2' comes from the fact that they are nearly equal and hence only half the applied voltage or half of local potential change is effective in raising the electron concentration while the other half is needed to increase the hole concentration. This can also be derived mathematically using the Boltzmann form of the nonequilibrium concentrations of the electrons and holes, (531.4C) and (531.6C). Following the discussion concerning the constancy of the majority carrier quasi-Fermi potentials $V_N(x)$ and $V_P(x)$ which led to $V_P - V_N \propto V$ in $-x_p<x<+x_n$, the nonequilibrium $P(x)N(x)$ product across the space-charge layer is then

$$P(x)N(x) = n_i^2\exp[q(V_P-V_N)/kT] \propto n_i^2\exp(qV/kT). \qquad (535.0)$$

(iii) As an approximation, the electron-hole recombination and generation rates can be assumed constant across the entire space-charge layer, $-x_p<x<x_n$, with an effective lifetime of τ_{pn}. The effective lifetime is approximately given by $\tau_{pn}=(\tau_{p0}+\tau_{n0})$ since a current pulse in the external circuit is completed only after an electron recombines with hole. The recombination event follows two steps: (a) capture of an electron by the recombination center followed by (b) capture of a hole by the recombination center. Thus, the lifetime is the sum of the two capture lifetimes.

The recombination current from the space-charge layer, I_{SNS}, can be obtained using these three keys and the second interpretation of the Shockley diode current. The second interpretation relies on recombination in an effective volume and was shown to be given by (electron charge)x(generation rate per unit volume)x(generation volume). Thus,

$$I_{SNS} = J_{SNS}\cdot(\text{Area})$$

$$= q[(n_i/\tau_{pn})\exp(qV/2kT)]\cdot(x_{sc})\cdot(\text{Area}) - I_R. \qquad (535.1)$$

$x_{pn}=(x_p+x_n)$ is the thickness of the space-charge layer. A is the junction area so that $x_{pn}A$ is the recombination volume. $-I_R$ is the reverse current due to generation of electrons and holes in the space-charge layer which we shall evaluate shortly. Dividing out the area, the current density is

$$J_{SNS} = [qn_i/(\tau_{p0}+\tau_{n0})](x_p+x_n)\exp(qV/2kT) - J_R \qquad (535.2)$$

where $J_R = I_R/A$ is the areal density of the space-charge-layer generation current density. It can be evaluated as follows. We note that the total space-charge layer current must vanish at zero applied voltage. Thus, letting $V=0$ and $J_{SNS}=0$ in (535.2), then

$$J_R = [qn_i/(\tau_{n0}+\tau_{p0})](x_p+x_n) = (qn_i/\tau_{pn})x_{pn} = J_2. \qquad (535.3)$$

Substituting this in (535.2), then the space-charge layer current is

$$J_{SNS} = J_2[\exp(qV/2kT) - 1]. \qquad \textbf{SNS Diode Equation} \qquad (535.4)$$

This current component has been termed the Sah-Noyce-Shockley (SNS) current by reviewers and researchers from Bell Telephone Laboratories in the 1960's. Other popular names were space-charge layer recombination-generation current, space-charge layer current, recombination current, and generation current. For ease of reference, we shall adopt the term, **space-charge layer current**, and the acronym SNS current following the tradition of named acronym.

The SNS diode equation (535.4) shows that the diode current at reverse applied voltages, $-J_R$, is voltage dependent and not like the Shockley diode current which saturates to a constant of $-J_1$. The voltage dependence comes from the voltage dependence of the space-charge layer thickness, $x_{pn}=x_p+x_n$, in J_2 given by (535.3). It increases as the square root of the reverse bias voltage as show in (531.1). The SNS diode equation also predicts an ideality factor in exp(qV/nkT) of $n=2$. A more detailed and accurate SNS theory shows that $1<n<2$ which agrees with many experimental diode data.

The space-charge-layer recombination-generation current given in (535.4) was derived using physical approximations. The error comes from the third key which assumed that the electron-hole recombination rate is a constant throughout the space-charge layer. This cannot be true at small forward bias voltages since the electron concentration drops from $n_i\exp(qV/2kT)$ at the center of the junction, x_j, to a lower value, $P_N\exp(qV/kT)$, at n-edge of the space-charge layer, x_n, as illustrated in Fig.531.2(c). Similarly, the hole concentration drops from $n_i\exp(qV/2kT)$ at x_j to $P_N\exp(qV/kT)$ at $-x_p$. Thus, the SNS diode equation (535.4) gives too high a recombination current density under forward bias. However, the approximation is good for reverse bias since the electron-hole concentrations are zero as illustrated in Fig.531.1(c) and the electron-hole generation rates in the space-charge layer are essentially constant and equal. The factor $[\exp(qV/2kT)-1]$ is also approximate since its derivation ignores (i) a difference or asymmetry in the spatial variation of the electron and hole concentration in the space-charge layer, $P(x)\neq N(x)$ and $P(x)\neq N(-x)$, and (ii) the asymmetry of the electron and hole recombination rates at the traps (or recombination centers) due to $\tau_{p0}\neq\tau_{n0}$ and $E_T\neq E_i$. The factor, [exp(qV/2kT)-1], is correct if we have a completely symmetrical trap and junction, with $\tau_{p0}=\tau_{n0}$, $E_T=E_i$, and $N_{DD}(x)=N_{AA}(-x)$.

The total diode current is the sum of the Shockley diode current given by (532.14) and the SNS diode current given by (535.4),

$$J = J_1[\exp(qV/kT) - 1] + J_2[\exp(qV/2kT) - 1]. \tag{535.5}$$

We now make some numerical estimates to illustrate the relative importance of space-charge layer current (SNS diode) to the quasi-neutral layer currents (Shockley diode). Taking the numerical values given in section 534 for the Shockley diode where $n_i=10^{10}cm^{-3}$, $\tau_{n0}=\tau_{p0}=10^{-6}s$ or $\tau_{pn}=\tau_{n0}+\tau_{p0}=2x10^{-6}s$, and assume $x_{pn}=10^{-4}cm$, then (533.4) has the magnitude of

$$J_2 = qn_ix_{pn}/\tau_{pn}$$

$$= 1.6x10^{-19}x10^{10}x10^{-4}/(2x10^{-6}) = 0.8x10^{-7}\ A/cm^2 \tag{535.6}$$

which is $0.8x10^{-7}/(4.8x10^{-10})=167$ times greater than J_1 given by (534.3C). Thus, recombination-generation of electrons and holes in the space-charge layer of silicon diodes (SNS diode) will dominate over that in the two quasi-neutral layers (Shockley diode) for reverse and small forward d.c. bias voltages since $J_2>>J_1$ for reverse bias and $J_2\exp(qV/2kT)>>J_1\exp(qV/kT)$ for small forward bias. At larger forward biases, $J_1\exp(qV/kT)$ becomes larger than $J_2\exp(qV/2kT)$ and the Shockley diode current dominates over the SNS diode current.

When the SNS current dominates at low forward voltage, few minority carriers are injected across the p/n junction boundary since much of the electrons and holes recombine with each other in the space-charge layer before they have a chance to get across the potential barrier of the space-charge layer. Thus, the Shockley current component would be very small since it comes from recombination of the injected minority carriers across the p/n boundary. Only at higher forward biases when the Shockley diode current dominates over the SNS diode current can the minority carriers be effectively injected across the p/n junction boundary. The voltage at which injection becomes effective can be estimated by equating the two current components, $J_1\exp(qV/kT) = J_2\exp(qV/2kT)$. This defines an **injection threshold voltage** given by

$$V_{IT} = (2kT/q)\cdot\log_e e(J_2/J_1)$$

$$= 2x0.025x\log_e(0.8x10^{-7}/4.8x10^{-10}) = 0.26V. \tag{535.7}$$

In summary, the following results and approximations are obtained for the current-voltage characteristic of a p/n junction diode.

$$J = J_{\rm Shockley} + J_{\rm SNS}$$

$$= J_1[\exp(qV/kT) - 1] + J_2[\exp(qV/2kT) - 1] \qquad (535.8)$$

$$J \approx J_{\rm SNS} = J_2[\exp(qV/2kT) - 1] \qquad \text{if } V \ll V_{IT} \qquad (535.8A)$$

$$J = 2J_{\rm SNS} = 2J_{\rm Shockley} \qquad \text{if } V = V_{IT} \qquad (535.8B)$$

$$J \approx J_{\rm Shockley} = J_1[\exp(qV/kT) - 1] \qquad \text{if } V \gg V_{IT}. \qquad (535.8C)$$

536 Breakdown of Reverse DC Current in a p/n Junction

In the introduction of this chapter, we briefly described the reverse current breakdown observed in a p/n junction diode. We shall now show in detail how the reverse current breakdown occurs in a p/n junction when it is biased at a large reverse bias voltage. The current breakdown behavior was illustrated in Fig.530.1(c). We noted that this current runaway or divergence at a definite reverse voltage comes about because of interband impact generation of electron-hole pairs by energetic electrons and holes. This generation process was illustrated in Fig.3613.1(b) and (d) and the process was labeled 13. This generation process has a large probability in the space-charge layer of a reverse biased junction because the large electric field in the space-charge layer under reverse bias accelerates many electrons to high kinetic energies. High kinetic energy is required to break the covalent (electron-pair) bond of 1.2 eV bond energy (Si energy gap, E_G) in order to create an electron-hole pair. This is illustrated in Fig.536.1(a).

We shall first use a simple model to show that current divergence can occur under reverse bias. Let us suppose that the probability for an electron to create an electron-hole pair after it passes through the SCL (space-charge layer) from the p-side to the n-side is I_N. Let I_P be the probability of generating an electron-hole pair by a hole crossing the SCL from n-side to p-side. Let us assume that $I_N = I_P = I$. Now, consider the case of one electron injected into the SCL from its boundary at $x = -x_n$. When this electron crosses the SCL, it will have generated I electron-hole pairs. The 1+I electrons will exit into the n-type quasi-neutral layer but the I holes just generated will cross the SCL to reach the p-side. These holes will generate IxI electron-hole pairs while passing through the SCL. Thus, continuing this enumeration and adding up all of the electrons generated, we have

$$1 + I + IxI + IxIxI + \ldots = \frac{1}{1 - I} = M_N. \qquad (536.1)$$

This is known as the **current or charge multiplication factor** and given the symbol, M_N (N=electron).

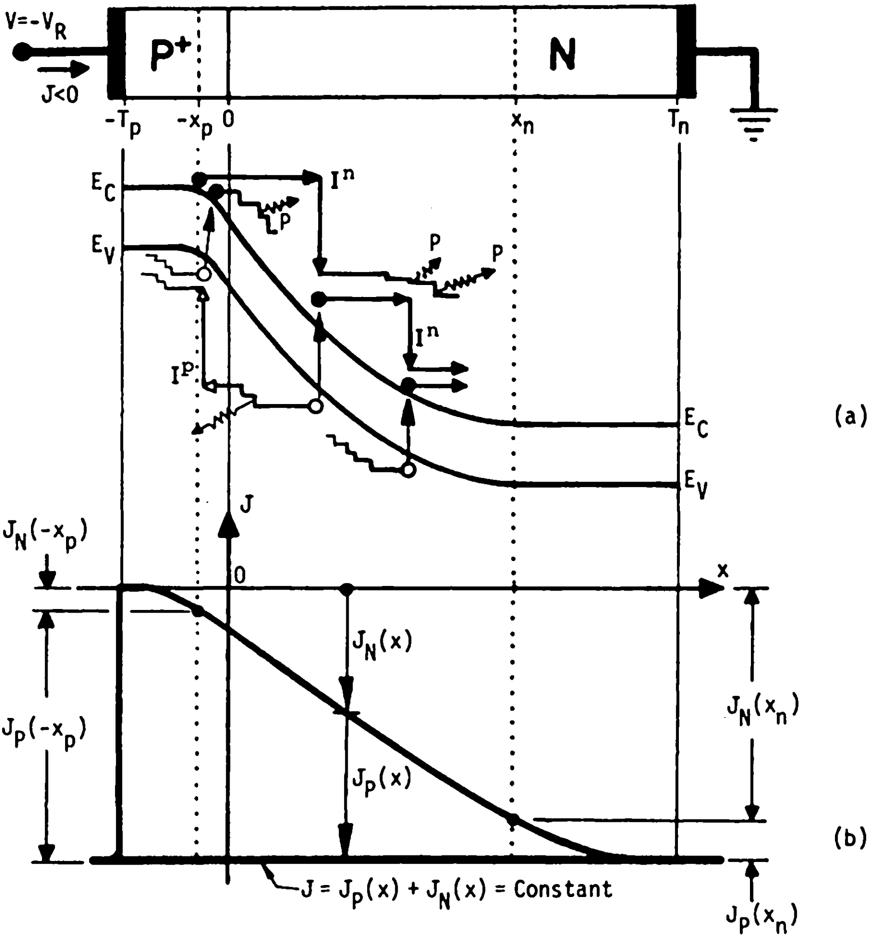

V=-V_R
J<0
P+
N
-T_p
-x_p
0
x_n
T_n
E_C
E_V
I^n
P
P
P
I^n
I^P
E_C
E_V
(a)
J
J_N(-x_p)
0
x
J_N(x)
J_p(-x_p)
J_N(x_n)
J_p(x)
(b)
J = J_p(x) + J_N(x) = Constant
J_p(x_n)

Fig.536.1 A p/n junction under large reverse bias to illustrate impact generation of electron-hole pairs in the space-charge layer by injected electrons from the p-side. (a) Energy band diagram. (b) Spatial dependences of the electron and hole current densities in the space-charge layer, $J_P(x)$ and $J_N(x)$. I = Interband impact generation of e-h pairs. P = Optical phonon emission.

A similar analysis will give a multiplication factor for the holes injected into the space-charge layer from the n-side. In general, the electron and hole impact generation probability, I_N and I_P, are not equal and the multiplication factors, M_N and M_P, have more complicated expressions than (536.1). In this simple example, we assume $I_P = I_N$. Thus, the multiplication factors for electrons and holes are also equal and given by,

$$M_P = M_N = M = 1/(1 - I). \tag{536.1A}$$

Since the p/n junction diode current is proportional to the electron and hole number or concentration, the current is increased by this multiplication factor and given by

$$\begin{aligned} J &= M(J_{Shockley} + J_{SNS}) \\ &= MJ_1[\exp(qV/kT) - 1] + MJ_2[\exp(qV/2kT) - 1]. \end{aligned} \tag{536.2}$$

Thus, it will diverge, when $M = \infty$ at $I = 1$. Physically, $I = 1$ means that one electron injected into the SCL from the p-side boundary will produce an infinite number of electron-hole pairs after the injected electron crosses the SCL. It also means that one hole injected into the SCL from the n-side boundary will produce an infinite number of electron-hole pairs after the injected hole crosses the SCL. Thus, the current breakdown condition or the breakdown voltage is the same whether an electron or a hole is injected into the space-charge layer to initiate the multiplication of the electron and hole numbers or concentrations.

In this simple analysis, we did not include the probability that an injected hole or electron can produce an electron-hole pair **inside** the SCL **before** it reaches the other boundary. These are illustrated in the energy band diagram given in Fig.536.1(a). We made the assumption that the electron will have a probability of I to generate an electron-hole pair **after** it crosses the entire SCL and reaches the far boundary ($x = x_n$). This is a good assumption since the electron will have the highest kinetic energy at the far boundary and hence the highest probability of generating an electron-hole pair by impact rupture of the covalent electron-pair bond at the far boundary.

Mathematical Formulation

The impact generation process and the current multiplication factor can be formulated and derived mathematically in order to include the e-h pairs generated inside the space-charge layer by impact which we just neglected. The mathematics involves rather simple differential equations and it delineates the important processes that control the magnitude of the impact multiplication factor. Thus, we will present this analysis. Again, we will assume that the electron and hole impact generation probabilities are equal. Since we now have the impact generation process, the continuity equation will contain this generation rate. The steady-state (d.c.) continuity equation for holes, from (350.2), was given in (532.1) which is repeated below

$$dJ_P/dx = q(G_P - R_P) + qG_{13n} + qG_{13p} \approx qG_{13n} + qG_{13p}. \tag{536.3}$$

Here, G_{13n} and G_{13p} are the electron-hole generation rates by impact generation caused by an energetic electron and hole respectively. The subscripts refer to the transition processes shown in Fig.3613.1(b) and 3613.1(d) respectively. From these two figures, it is evident that

$$G_{13n} = \alpha_n \theta_n N \quad \text{and} \quad G_{13p} = \alpha_p \theta_p P \qquad (536.4)$$

where $\alpha_n\theta_n$ and $\alpha_p\theta_p$ are proportionality coefficients which are separated into two parts: α and θ.

α_n and α_p are the **impact ionization coefficients** or interband impact generation probabilities (or probability coefficients) of electron-hole pairs by electron and hole impact respectively. The term, impact ionization, was borrowed by K. G. MacKay and K. B. MacAfee of Bell Telephone Laboratories in 1951 from the theory of impact ionization of gas due to electrons hitting atoms and molecules such as those occurring in gas-filled tubes of neon signs and fluorescent light during their investigation of the mechanism of current divergence in silicon p/n junction diodes. It is a misnomer when used in semiconductors since the semiconductor is not ionized because electron-hole pairs are generated and the change of space-charge is zero. A better term with correct physics is '**interband impact generation probability coefficient**' which can be abbreviated to **interband impact coefficient** without a mix-up in physics. We cannot call it probability since $\alpha_n \Delta x$ is dimensionless and is a normalized probability while α_n has the dimension of cm^{-1} and is not normalized to unity. We shall use the term, **impact generation coefficient**, interband impact coefficient, and occasionally, impact ionization coefficient, when referring to the original papers of MacKay and MacAfee and later papers by van Overstraeten and Moll which gave accurate data taken on specially designed Si diodes.

θ_n and θ_p are the drift velocities of electrons and holes in the high electric field of the SCL of the reverse biased p/n junction. They are the product of the drift mobility and the electric field, $|\mu_n E|$ and $|\mu_p E|$. They each saturate to a constant value at a high electric field, about 1.0×10^7 cm/s at $(0.75\rightarrow2)\times10^4$ V/cm for electrons and 0.8×10^6 cm/s at $(2\text{-}8)\times10^4$ V/cm for holes in Si as indicated in section 314 and Fig.314.1. The higher field values are at full saturation of velocity. The saturated velocity values are determined by the kinetic energy losses by electrons and holes via optical phonon emissions during lattice scattering.

It is intuitively clear that the interband impact generation rates in pair/cm^3-s, denoted by G_{13n} and G_{13p}, should be proportional to the carrier concentration, N or P, as suggested by Fig.3613.1(b) and (d). In the approximation made in (536.3) above, we assumed that the impact generation of electron-hole pairs is the dominant generation process. We have shown that in the space-charge layer, electron-hole generation-recombination-trapping via the thermal or Shockley-Read-Hall (SRH) trap-band transitions will dominate the d.c. current at reverse and at small forward biases. Thus, a more accurate and general analysis must also include the SRH generation rate, G_{21} in (536.3). But the arithmetic and integration are more complicated which could obscure the fundamentals. In order to delineate and focus on the physics, we shall neglect the SRH thermal generation of electron-hole pairs in the space-charge layer represented by G_{21} and will only consider the interband impact generation process in this example.

The current in the two quasi-neutral layers is, dominated by minority carrier diffusion which was used to derive the Shockley diode equation. The current flowing through the space-charge layer, however, is dominated by drift due to the high electric field and very low carrier density (depletion condition used earlier). Thus, the hole current in the space-charge layer is

$$J_P = q\mu_p EP - qD(dP/dx) \approx q\mu_p EP = -q\theta_p P \qquad (536.5)$$

where we have used the relationship between the drift velocity and mobility, $\theta_p = \mu_p(-E)$. The negative sign comes in because E is negative in the coordinate system we are using and the drift velocity is defined as a magnitude without sign. A similar set of two equations for the electron current J_N can be obtained.

Summarizing the results just obtained, we have the following solutions in the space-charge layer spanning $-x_p \le x \le +x_n$.

$$dJ_P/dx = qG_{13n} + qG_{13p} = q\alpha_n\theta_n N + q\alpha_p\theta_p P \qquad (536.6)$$

$$J_P = -q\theta_p P \qquad (536.7)$$

$$J_N = -q\theta_n N \qquad (536.8)$$

and

$$dJ_N/dx = -q\alpha_p\theta_p P - q\alpha_n\theta_n N. \qquad (536.9)$$

Summing (536.6) and (536.9) gives

$$dJ_P/dx + dJ_N/dx = 0 \qquad (536.10)$$

or

$$J = J_P(x) + J_N(x) = \textbf{spatially constant.} \qquad (536.10A)$$

The spatial constancy of the one-dimensional current density, J, in the space-charge layer given by (536.10A) is just a mathematical statement of Kirchoff's current law. It is spatially constant since there are no generation-recombination-trapping losses of electrons and holes in the space-charge layer in our model. In the model, we neglected the thermal (SRH) process and all other generation-recombination-trapping processes at the traps in the space-charge layer. Only the interband impact generation process is retained to simplify the arithmetic so as to delineate and focus on the physics.

Basic Physics of the Parameters

The impact generation (ionization) coefficients by electrons, α_n, and holes, α_p, are functions of the electric field. Their magnitudes depend on the energy gap, electron and hole effective masses as well as the interaction of the electrons and holes with lattice vibrations or phonons via lattice scattering.

The strong dependence on the electric field comes about because the electrons and holes must be accelerated to a kinetic energy exceeding the threshold energy necessary to break a covalent bond in order to generate an electron-hole pair. In section 3613, the interband impact threshold energy was shown to be $3E_G/2 \approx 1.8$eV for an idealized Si energy band (spherical E-k conduction and valence bands and equal effective masses $m_e = m_h$).

The phonon dependence comes about since the electrons and holes are scattered by the longitudinal acoustic and optical phonons or lattice vibrations in the space-charge layer, just like the scattering they experience in a field-free homogeneous semiconductor (section 313) which determines the temperature dependence of the drift mobilities. The elastic or nearly elastic collisions from acoustic phonon scattering randomize the velocity or the direction of travel which will only increase the transit time of electrons through the space-charge layer. However, the inelastic collisions involving optical phonon emission will significantly reduce the probability for the electrons and holes to reach the impact threshold kinetic energy required to break a bond while passing through the space-charge layer. Pictorially, longitudinal optical phonon emission is the process of setting off an outgoing or departing longitudinal lattice vibration wave at the point of collision of an energetic electron with a vibrating host atom. It is more important than the phonon-absorbing electron scattering events at high electron kinetic energies for two reasons: (i) there are few phonons of high energies at room temperatures to be absorbed by the electrons and (ii) the probability of phonon emission increases rapidly once the kinetic energy of electrons and holes exceed the optical phonon energies which are about 25 and 60 meV in Si. The reason for limiting to the longitudinal phonons and not including transverse phonons is the same as that given for drift mobility: momentum or k-vector conservation selects the longitudinal waves of lattice vibrations. It is also obvious that there is a very large probability for the electron to emit several phonons while being accelerated to the interband impact generation threshold energy because of the large difference between the phonon and impact threshold energies. Taking the optical phonon energy of 60 meV and the impact threshold energy of $3E_G/2 = 1.8$eV for the ideal Si energy bands, we get an energy ratio of 1.8eV/60meV = 1.8/0.06 = 30 optical phonons. So, if the optical phonon scattering probability is 3.33% during the acceleration through the space-charge layer to gain the 1.8 eV energy, the electron would generate or set off one 60 meV optical phonon.

The phonon emission processes are graphically illustrated in Fig.536.1(a) as resistance-like arrows and labeled by P. The interband impact generation of electron-hole pairs are represented as upward vertical arrows, labeled by I^n or I^p for impact by energetic electron or hole. Because phonon emission dominates while the temperature-sensitive phonon absorption is unimportant, interband impact generation rate is insensitive to temperature. Thus, the p/n junction breakdown voltage has a rather weak temperature dependence.

The electric field dependence can be approximated by the following exponential formula, borrowed from gaseous electronics by MacKay and MacAfee of the Bell Telephone Laboratories in 1951,

$$\alpha = \alpha(E) = \alpha_0\exp(-b/E). \tag{536.11}$$

E is the magnitude of the electric field in the junction. It is position dependent [See Fig.531.1(b)], $E=E(x)$, so that α is also position dependent, $\alpha=\alpha(x)$.

Physically, α_0 is the maximum number of optical phonons that can be generated when an energetic electron or hole passes through Δx. But, this is reduced by exp(-b/E) since some of the electrons and holes will suffer energy-loss (inelastic) collisions with the lattice vibration via optical phonon emission on their way to Δx. These physics lead to the following simple derivation of (536.11).

(1)
$$\begin{aligned}\alpha_0\Delta x &= \text{Maximum number of phonons that can be generated in } \Delta x\\ &= (\text{Potential energy gained across } \Delta x)/(\text{phonon energy})\\ &= (qE\Delta x)/(\hbar\omega_0)\end{aligned}$$

so

$$\alpha_0 = qE/(\hbar\omega_0) \tag{536.11A}$$

(2)
$$\begin{aligned}b/E &= (\text{distance traveled to reach impact threshold})\\ &\quad\ \text{divided by (mean free path)}\\ &= (E_I'/qE)/\lambda_0\end{aligned}$$

or

$$\begin{aligned}b &= E_I'/q\lambda_0\\ &= E_I/\lambda_0.\end{aligned} \tag{536.11B}$$

E is the magnitude of the electric field. q is the magnitude of electron charge. $\hbar\omega_0$ is the longitudinal optical phonon energy. E_I' is the impact threshold energy in Joules and E_I is that in eV. λ_0 is the mean-free-path of energetic electrons or holes in the presence of optical phonon scattering. In Si, it is about 50A for energetic holes and 75A for energetic electrons, giving $b=3.6\times10^6$ V/cm and 2.4×10^6 V/cm respectively. The above simple physical derivation and physics-focused tutorial description follow a slightly more elaborate but also physical derivation given by Shockley at Stanford in 1961 when he proposed and demonstrated this simple **ballistic or lucky electron model** to explain the electric-field dependence of the α_n and α_p data. His model and analyses led to detailed and accurate measurements presented by van Overstraeten in his Ph.D. thesis under Moll at Stanford in 1963 and van Overstraeten's further and elaborate measurements with his Ph.D. students at the Katholiek Universteit Leuven, Belgium in 1970.

Simple Solution

To find a simple solution of (536.6) to (536.10), we assume that $\alpha_n = \alpha_p = \alpha$, then we get

$$dJ_P/dx = -\alpha_n J_N - \alpha_p J_P = -\alpha(J_N+J_P) = -\alpha J. \qquad (536.12)$$

The assumption of $\alpha_n=\alpha_p=\alpha$ gives the simple result above and makes the solution of the differential equation trivial which was the reason for making this assumption. Thus, integrating (536.12) we have

$$J_P(x) - J_P(-x_p) = -\int_{-x_p}^{x} \alpha(x) J dx = -J\int_{-x_p}^{x} \alpha(x)dx. \qquad (536.13)$$

Using the spatial dependences of $J_P(x)$ and $J_N(x)$ illustrated in Fig.536.1(b), then the Kirchoff's current law of (536.10A) gives

$$J_N(x_n) = J - J_P(x_n) = J + J_{p0}, \qquad (536.14)$$

and

$$J_P(-x_p) = J - J_N(-x_p) = J + J_{n0}. \qquad (536.15)$$

We used the boundary condition $J_P(x_n)=-J_{p0}$ in (535.14) and $J_N(-x_p)=-J_{n0}$ in (536.15) to give the results there. These boundary conditions come from the d.c. solutions of the diffusion-recombination-generation current which gave us the Shockley ideal diode equation (532.14).

Evaluating (536.13) at $x=x_n$ and using these boundary conditions, then

$$J = -(J_{p0}+J_{n0})\left[1 - \int_{-x_p}^{x_n} \alpha(x)dx\right]^{-1} = -\frac{J_1}{1 - I} = -MJ_1. \qquad (536.16)$$

This shows that the impact ionization probability, I, is related to the fundamental parameter, the interband e-h pair generation coefficient, α. It is the integrated α over the space-charge layer:

$$I = \int_{-x_p}^{+x_n} \alpha(x)dx = \int_{-x_p}^{+x_n} \alpha_0 \exp[-b/|E(x)|]dx. \qquad (536.17)$$

The current multiplication factor, $M=1/(1-I)$, of a p/n junction diode with a known $N_{DD}(x)-N_{AA}(x)$ can be computed as a function of applied reverse dc voltage as follows. The parameters α_0 and b are given from previous experiments. E(x) vs x is determined first from the given $N_{DD}(x)-N_{AA}(x)$ by solving the Poisson equation numerically or analytically in the depletion approximation at an applied dc voltage. Next, (536.17) is integrated numerically to give I which is then used to compute M using $M=(1-I)^{-1}$. This is repeated at each applied dc voltage to give a curve of I vs V_R and M vs V_R. The breakdown voltage is then the value of V_R when $I=1$ and $M=\infty$. For example, in the abrupt n+/p junction with the space-charge layer on the p-side entirely (i.e. $N_{DD}>>N_{AA}$), I is given by

$$I = \int_0^{x_{pn}} \alpha_0 \cdot dx \cdot \exp\left[-b/\{E_M[1-(x/x_{pn})^2]\}\right] \qquad (536.18)$$

where E_M and x_{pn} are given by (531.2A) and (531.1) respectively with $N_M=N_{AA}$. This integral can be evaluated numerically.

Figure 536.2(a) gives the computed breakdown voltages of abrupt silicon p+/n or n+/p junctions as a function of N_{AA} or N_{DD} assuming that N_{AA} and N_{DD} are spatially constant and the entire space-charge layer is on the lower doped side of the junction. Figure 536.2(b) gives the breakdown voltages of these junctions as a function of the resistivity of the lower-doped side of the asymmetrical junctions. The resistivity is calculated from the given N_{AA} or N_{DD} using the mobility data in Si at 297K given in Fig.313.5. These breakdown voltage curves give the theoretical minimum values based on the given set of α_n and α_p data for Si at a given dopant impurity concentration or resistivity. It is the minimum since a practical p/n junction manufactured by high temperature diffusion will have a lower electric field in the space-charge layer. This is due to the gradual impurity concentration profile compared with the electric field of the assumed abrupt n+/p impurity profile in the theoretical breakdown voltage calculations. The lower electric field would require a higher voltage to reach the current divergence condition, $I=1$ and $M=\infty$. However, the breakdown voltage in practical p/n junction diodes can also be lower than the theoretical minimum given by these figures due to the high electric fields localized at (i) perimeter of the p/n junction diode, and (ii) imperfections such as impurity and defect clusters which intercept the junction space-charge layer. The high electric field at these imperfections will cause $I=1$ and $M=\infty$ to occur at these imperfection regions, at a lower applied voltage than theory.

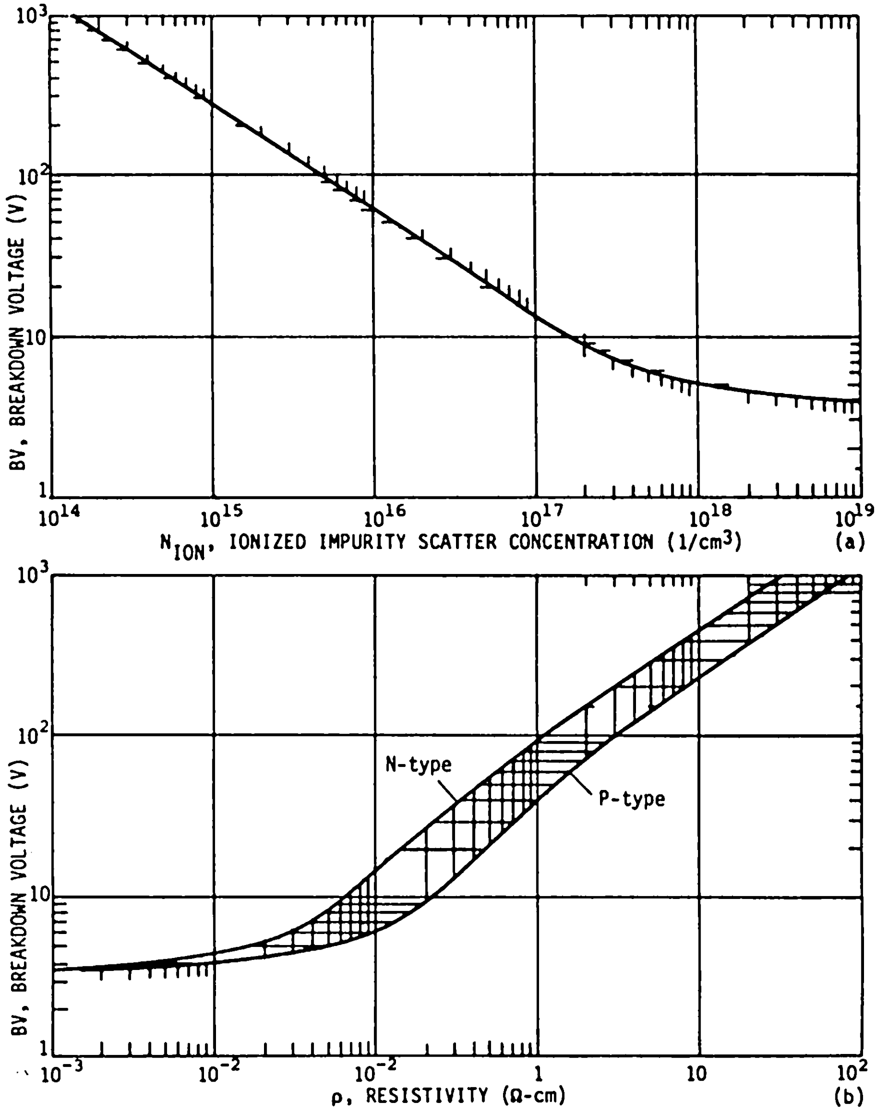


Fig.536.2 The breakdown voltage of Si asymmetrical p+/n and n+/p junctions at 300K as a function of (a) ionized impurity concentration, $N_{ION}=N_{DD}$ for p+/n and $N_{ION}=N_{AA}$ for n+/p junctions, and (b) resistivity.

537 Experimental-Theoretical Comparison

The d.c. theory we have just developed for p/n junction diodes compares well with experimental data. Figures 537.1(a) and (b) show the first historical comparison made in 1956 between theory and experiment given by Sah, Noyce, and Shockley. [Proc. IRE, 45, 1228, Sept.1957].

1957 Sah, Noyce and Shockley: Carrier Generation and Recombination in P-N Junctions 1237

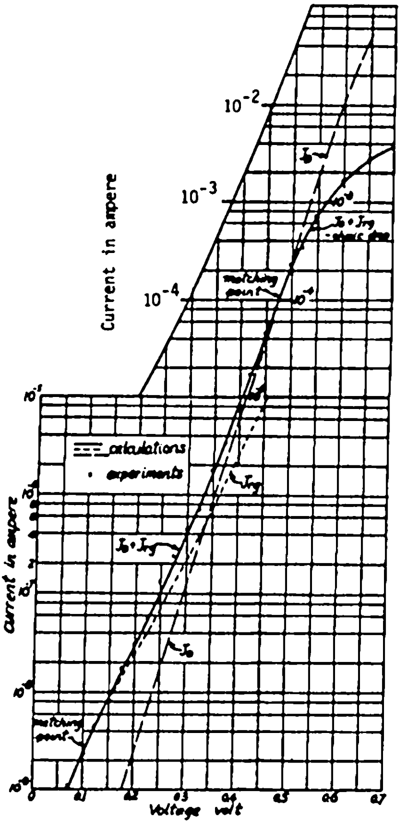


Fig. 11—Forward characteristics of a silicon p-n junction.

NOTE: Notation Changes.

1957 Sah, Noyce, Shockley Paper	1989 EEL3396 Sah's Book Here
J_D	$J_{Shockley}$
J_{rg}	J_{SNS}

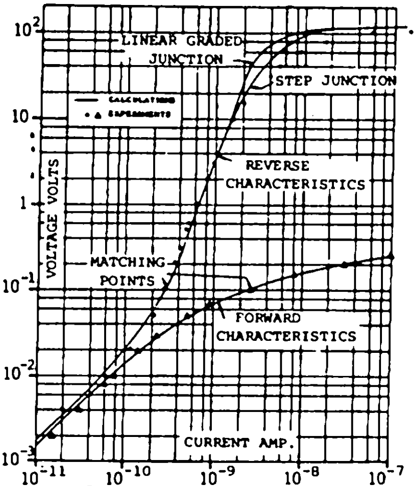


Fig. 12—Silicon p-n junction characteristics.

Fig.537.1 One of the first historical comparison of experiments with theory of the d.c. current-voltage characteristics of an Si diffused p+/n junction given by Sah, Noyce and Shockley in 1956. The theory was computed using a three-point match indicated on the figure and discussed in the text. Some of the original handwritten labels are replaced by larger typewritten labels for legibility. From Proc.IRE, 45, 1228, Sept.1957. The data was replotted in Fig.500.1(c).

The silicon p+/n diode has an area of 0.62mm^2. The n-type substrate has $N_{DD}=2.2\times10^{15}$ cm^{-3}. The junction is diffused with a boron surface concentration of $N_{AA}(x=0)=10^{21}$ cm^{-3} and a junction depth of 11.8μm. Since the lifetimes were not measured, they were used to match the theory to the data at two points. A third matching point was used to give the energy level of the recombination centers, E_T. E_T was needed to compute the reverse current due to e-h generation in the space-charge layer. The breakdown voltage, 170V, was also computed by setting the integral $I=1$. Experimental BV was about 120V. Breakdown due to high electric field at impurity cluster and physical defect sites discussed in sections 530 and 536 were the causes of the lower breakdown voltage in this diode since the diode was a mesa diode whose cross section is shown in Fig.540.1(a) and it did not have high-field junction curvature of the planar diffused junctions made by impurity diffusion through holes in the oxide such as the oxide-protected planar Si p/n junction diode shown in Fig.540.1(a') and the junctions of the planar MOS and bipolar transistors shown in Figs.610.3 and 710.3. This was verified by the large number of light emitting spots seen under a microscope when the 1956 mesa diode was reverse biased to very high breakdown current using a Tektronix transistor d.c. curve tracer. The theoretical I-V curve was obtained by 3-point match which gave the empirical or curve-fitting values of $\tau_{p0}=0.012\mu$s, $\tau_{n0}=4.3\mu$s, and $E_T-E_I=0.119$eV. Such an empirical fit is known today as **parameter extraction** in modern VLSI circuit design methodology.

Data of a later (1962) silicon diffused and oxide-passivated diode are shown in Figs.537.2(a) and (b) whose cross section is shown in Fig.540.1(a'). This was fabricated using the advanced silicon photolithographic and oxide-passivation technologies comparable to those used today. Consequently, the electron and hole lifetimes were high, $>10\mu$s, and the reverse current was low. The breakdown voltage was still lower than the theory. This was due to high electric fields at the junction curvature and the oxide-protected surface junction shown in Fig.540.1(a') rather than the bulk impurity clusters in the 1956 diodes.

The voltage dependences of the reverse current-voltage characteristics at many temperatures are shown in Fig.537.2(a) and they verify the theories just discussed. At each high temperatures, Fig.537.2(a) shows a saturation of reverse current as the reverse voltage increases. This is the indicator of the dominance of the Shockley diode component in the higher temperature range since the Shockley diode equation, (532.14), predicts reverse current saturation to a constant value, $-J_1$ given by (533.2) and (533.4), when the reverse voltage is increased.

The experimental temperature dependences of the reverse current are shown in Fig.537.2(b) at -1 volt reverse bias. They agree with the temperature dependence of the two theoretical reverse current components: the Shockley diode current due to carrier generation in the two quasi-neutral layers, and the SNS diode current due to carrier generation in the junction space-charge layer. From (532.15) for J_1 and (535.3) for J_2, their theoretical temperature dependences are:

$$
\begin{aligned}
J_1 &= q(D_pP_N/L_p) \quad + q(D_nN_P/L_n)\\
&= q(\sqrt{D_p/\tau_{p0}})P_N + q(\sqrt{D_n/\tau_{n0}})N_P\\
&\approx q[\sqrt{D_p/\tau_{p0}}) + \sqrt{D_n/\tau_{n0}})][(n_i^2/N_{DD}) + (n_i^2/N_{AA})]\\
&= f_1(T)n_i^2 \propto n_i^2 \propto \exp(-E_G/kT) \quad \text{Shockley Current} \qquad (537.1)
\end{aligned}
$$

and

$$J_2 = f_2(T)n_i \propto n_i \propto \exp(-E_G/2kT) \quad \text{SNS Current.} \qquad (537.2)$$

The predominant temperature dependence comes from $n_i^2 \propto \exp(-E_G/kT)$ for the Shockley current, and from $n_i \exp(-E_G/2kT)$ for the SNS current. The pre-exponential factors $f_1(T)$ and $f_2(T)$ vary with temperature much slower than $\exp(-E_G/kT)$ or $\exp(-E_G/2kT)$. $f_1(T)$ comes from the temperature dependence of the diffusivities, D_p and D_n, or mobilities, $D_p=(kT/q)\mu_p$ and $D_n=(kT/q)\mu_n$, and a possible small temperature dependence of the lifetimes, τ_{p0} and τ_{n0}. $f_2(T)$ comes from the temperature dependence of $\tau_{pn}=\tau_{p0}+\tau_{n0}$ and x_{pn} via $V_{bi}(T)$. These temperature dependences follow power laws, resulting in $f_{1,2}(T) \propto T^k$, where k not likely to be outside of the range $-5<k<5$.

The term, **thermally activated process**, is used when a device characteristic varies exponentially with temperature as $\exp(-\Delta E/kT)$. ΔE is known as the **thermal activation energy**. It is analogous to the thermal activation energy of a chemical reaction where the term originated. Thus, the Shockley current has a thermal activation energy of the **full energy gap**, $\Delta E=E_G$, while the SNS current has a thermal activation only **one-half of the energy gap**, $\Delta E=E_G/2$.

Due to the strong temperature dependence of $\exp(-\Delta E/kT)$, when $\log_e$(rate) vs $1/T$ is plotted, a straight line segment results with a negative slope of $\Delta E/k$. Several line segments of different slopes in an experimental data would indicate several generation species and locations. This **thermal activation plot** is known as the **Arrhenius plot**. The experimental data of the reverse current at -1 vol are plotted in Fig. 537.2(b). It has two slopes: E_G at higher temperatures when the Shockley current dominates and $E_G/2$ at lower temperatures when the SNS current dominates.

The ratio, J_2/J_1, gives relative importance of the SNS and Shockley currents. Consider a p+n diode and let $\tau_{p0}=\tau_{n0}$, then

$$
\begin{aligned}
J_2/J_1 &= (qn_i/\tau_{pn})x_{pn} \div [q\sqrt{D_p/\tau_{p0}}(n_i^2/N_{DD})]\\
&= (N_{DD}/n_i)(x_{pn}/2\sqrt{D_p\tau_{p0}}) \qquad (537.3)
\end{aligned}
$$

Thus, SNS current is important at low resistivity (high N_{DD}), low temperature and large energy gap (low n_i), and low lifetime. Si and compound semiconductor have larger SNS current than Shockley current due to their large energy gaps.

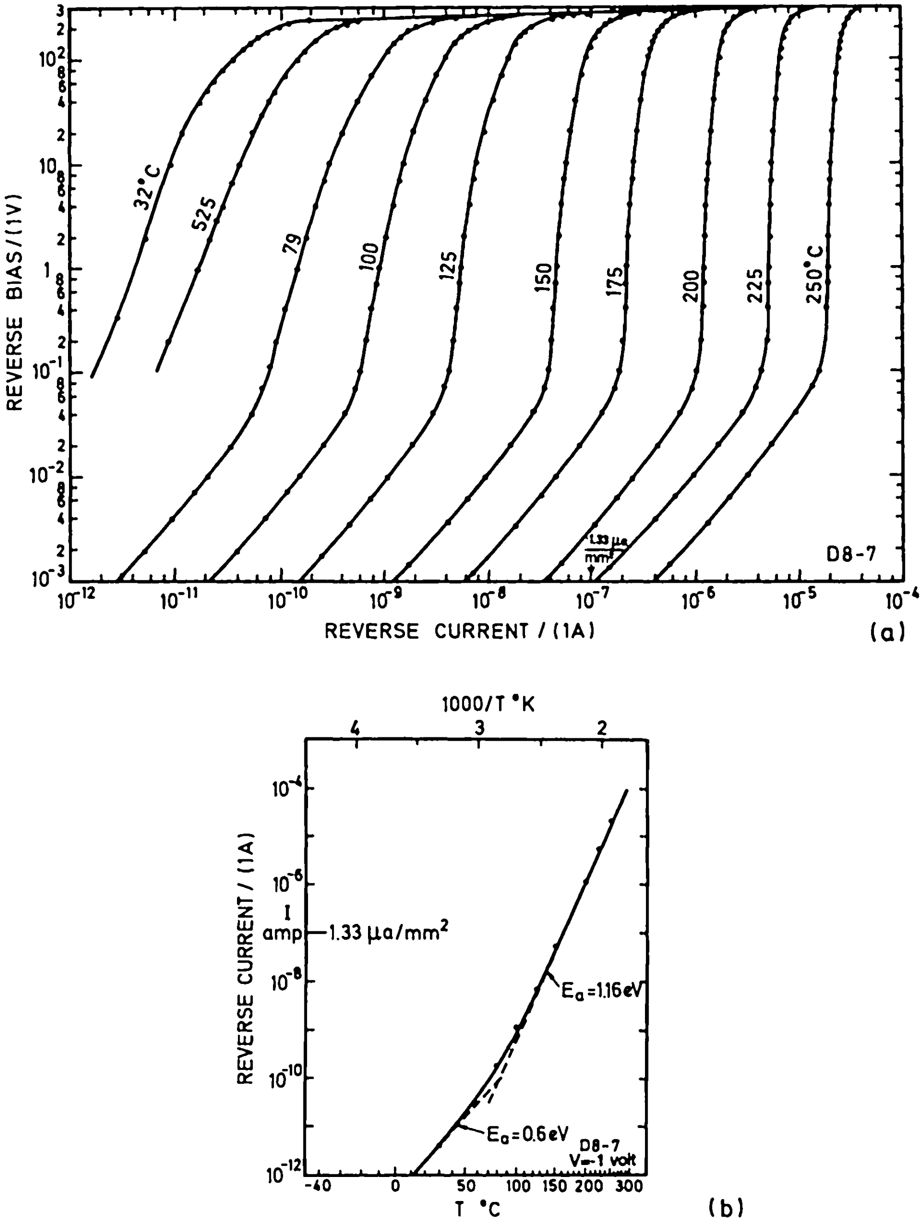
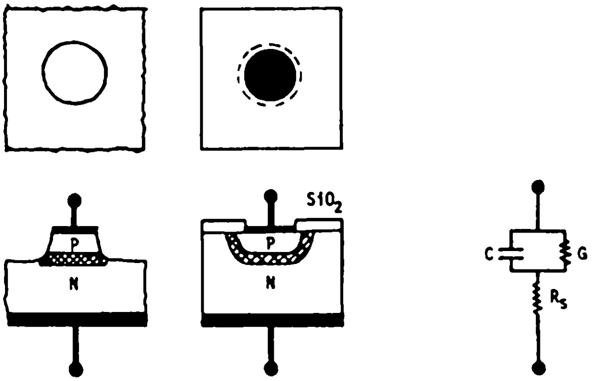


Fig.537.2 **(a)** The d.c. reverse current-voltage characteristics of an experimental diffused and oxide passivated silicon p/n junction diode with diode temperature as the parameter. **(b)** The thermal activation plot of the reverse current at a constant reverse voltage as a function of 1/T.

540 SMALL-SIGNAL CHARACTERISTICS OF A P/N JUNCTION

The simplest small-signal equivalent circuit of an ideal p/n junction diode is just a capacitance **in parallel** with a conductance shown in Fig.540.1(b). The top and cross-sectional views of the 1957 Si mesa diode and the 1960 Si planar diode are shown in Fig.540.1(a) where the space-charge layer is cross-hatched. The conductance G (or resistance) originates from the recombination-generation losses of electrons and holes at the traps in the three layers of the junction, the two quasi-neutral layers and the space-charge layer. The capacitance C represents the **mobile** charges (electrons in the conduction band and holes in the valence band) stored in the three layers. A capacitor representing the **immobile** charges stored on the traps or generation-recombination-trapping (g-r-t) centers, C_t, is neglected in this elementary treatment since it is usually very small due to the low trap concentration. These circuit elements are functions of the d.c. voltage applied to the diode. At high d.c. forward bias voltages and currents and at high frequencies, four small series resistances should also be added to represent the **bulk resistances** of the two quasi-neutral layers, R_{pb} and R_{nb}, and the **contact resistances**, R_{pc} and R_{nc}, of the two external leads making contact to the p-type and n-type surfaces of the semiconductor wafer. The bulk series resistances represent the majority carrier energy losses due to electron-phonon and hole-phonon scattering in the two quasi-neutral layers. As described in chapter 2, they are the series resistances to the drift current. The contact resistances represent the energy loss due to carrier recombination-generation in the interfacial layer of two metal/semiconductor junctions at the two surfaces of the semiconductor diode wafer. The two bulk and two contact series resistances are combined and represented by R_s which is given by $R_s = R_{pc} + R_{nc} + R_{pb} + R_{nb}$.



1957 Si Mesa p/n Junction Diode (a)   1960 Si Planar p/n Junction Diode (a')   Small-Signal Equivalent Circuit (b)

Fig.540.1(a) The physical structure and (b) the small-signal equivalent circuit of a p/n diode.

541 Small-Signal Charge-Control Circuit Elements

The small-signal equivalent circuit elements, G and C, of the junction can be computed using the differential of current and charge when a small change occurs in the applied voltage, $G=dJ/dV$ and $C=dQ/dV$. This is known as the charge control analysis method and the result is known as the charge control model. This method is simple to apply and the result is simple to use. However, the result is only approximately correct since it does not take into account three sources of signal delay: (i) the diffusion-drift delay through the two quasi-neutral base layers of thicknesses T_p and T_n; (ii) the drift-diffusion delay through the high-field space-charge layer of thickness $x_{pn}=x_p+x_n$; and (iii) the trapping delay due to electron and hole capture and emission at the traps or g-r-t centers in the three layers. But simplicity in derivation and concept makes the charge control model extremely popular. The neglect of trapping delay represented by a charge storage capacitance of the traps is a good approximation since the concentration of the traps in production diodes and transistors in integrated circuit chips is usually very low compared with the concentration of the electrons, holes, dopant acceptors and dopant donors. Only in two device types, the p/n junction power rectifier diodes and the p/n/p/n silicon controlled rectifier (SCR) power switches used in 60-cycle (60-Hertz) power control circuits, is the trap or g-r-t center concentration high. But, their switching applications do not require a knowledge of the small-signal circuit model. The switching models, waveforms, and delay time calculations will be discussed later, in sections 55n.

The conductance element is given by dJ/dV which can be computed from (532.13) and (535.2). The conductance of the three layers are given by

$$G_p = (qP_N L_P/\tau_p)(q/kT)\exp(qV/kT) = (qJ_{p0}/kT)\exp(qV/kT) \qquad (541.1)$$

$$G_n = (qN_P L_n/\tau_n)(q/kT)\exp(qV/kT) = (qJ_{n0}/kT)\exp(qV/kT) \qquad (541.2)$$

$$\begin{aligned} G_{pn} = {} & (qn_i x_{pn}/\tau_{pn})(q/2kT)\exp(qV/2kT) \\ & + (qn_i/\tau_{pn})(dx_{pn}/dV)[\exp(qV/2kT) - 1]. \end{aligned} \qquad (541.3)$$

The total junction conductance shown in Fig.540.1(b) is the sum:

$$G = G_p + G_{pn} + G_n. \qquad (541.4)$$

The capacitance element contains also three parts, one from each of the three layers. They are computed using the charge control method, $C=dQ/dV$, where dQ is the change of the **stored mobile charges** in the three layers (either hole or electron but not both, why?). The electron and hole concentrations at $V=0$ and $V=+\Delta V>0$ are shown in Fig.541.1(a) and the logarithmic of the differences, ΔP and ΔN, in Fig.541.1(b). To illustrate the differences clearly, $\Delta V=4.6kT/q$ is used to give $\exp(q\Delta V/kT)=100$.

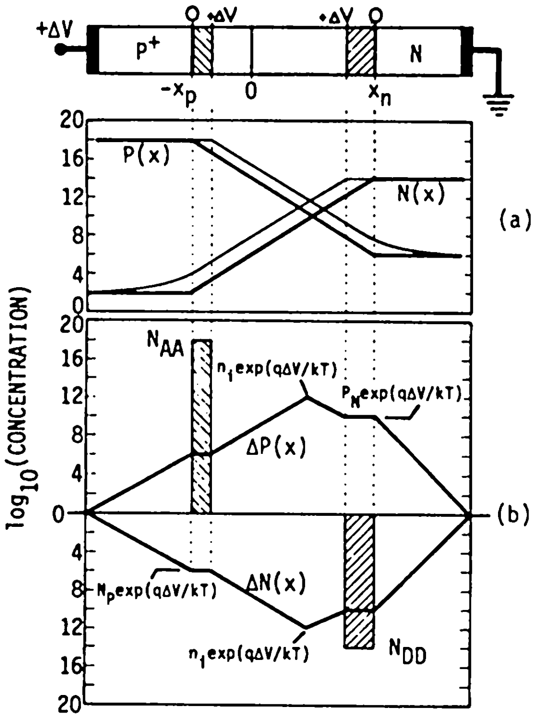

+ΔV
0
+ΔV
+ΔV
0
p+
N
-xp
0
xn
20
16
12
8
4
0
P(x)
N(x)
(a)
log10(CONCENTRATION)
20
16
12
8
4
0
4
8
12
16
20
NAA
ni exp(qΔV/kT)
PN exp(qΔV/kT)
ΔP(x)
(b)
Np exp(qΔV/kT)
ΔN(x)
ni exp(qΔV/kT)
NDD

Fig.541.1(a) The carrier concentrations at at V=0 (dark lines) and V=+ΔV=4.6kT/q (light lines), and (b) the change of the carrier concentrations between V=+ΔV and V=0.

Using P(x) of (532.10), the stored excess minority carrier (holes) charges per unit diode area in the n-type quasi-neutral layers is given by

$$Q_P = q\int_{x_n}^{T_n{}^{+\infty}} [P(x)-P_N]dx = qP_NL_p[\exp(qV/kT) - 1]. \qquad (541.5)$$

Assuming $P(x) \propto N(x) = n_i \exp[\exp(qV/2kT)-1]$ in $-x_p < x < +x_n$, the approximate d.c. density of hole or electron charges stored in the space-charge layer is then

$$\begin{aligned} Q_{PN} = q\int_{-x_p}^{x_n} P(x)dx &= qn_i(x_n+x_p)[\exp(qV/2kT) - 1] \\ &= qn_i \cdot x_{pn} \cdot [\exp(qV/2kT) - 1]. \end{aligned} \qquad (541.6)$$

Using a integration similar to that leading to (541.5), the minority carrier (electron) charges stored in the p-type quasi-neutral layer can be written down. Without algebra, instead, interchanging n and p in (541.5) the answer is

$$Q_N = qN_PL_n[\exp(qV/kT) - 1]. \qquad (541.7)$$

The capacitances due to the minority carrier charges stored in these three layers are then

$$C_P = dQ_P/dV = (q^2/kT)P_N L_P \exp(qV/kT) = G_P\tau_P \qquad (541.8)$$

$$C_n = dQ_N/dV = (q^2/kT)N_P L_n \exp(qV/kT) = G_n\tau_p \qquad (541.9)$$

$$C_{pn} = dQ_{PN}/dV = (q^2/2kT)n_i x_{pn}\exp(qV/2kT) + qn_i(dx_{pn}/dV)[\exp(qV/2kT)-1] = G_{pn}\tau_{pn} \qquad (541.10)$$

In addition, there is a capacitance from adding majority carrier charges to the edges of the space-charge layer due to $+\Delta V$, shown as by the two shaded rectangles in Fig.541.1(b). This alters the amount of majority carrier charges depleted from the space-charge layer, dQ_D. It is evident from the figure that the capacitance is given by the parallel-plate capacitance

$$C_d = |dQ_D/dV| = \varepsilon_s/x_{pn}. \qquad (541.11)$$

The total capacitance is the sum of these four terms:

$$C = C_p + C_d + C_{pn} + C_n. \qquad (541.12)$$

The capacitances due to the majority carrier mobile charge densities stored in the two quasi-neutral layers need not be calculated. The reason is that they are already included because the changes of the majority and minority carrier densities in the two quasi-neutral layers due to a change of the applied voltage ΔV are equal when the trap concentration is small as we have assumed, i.e. $\Delta p = \Delta n + \Delta n_T \simeq \Delta n$ if $\Delta n_T << \Delta n$. If the trap concentration is not small, Δn_T cannot be neglected and the majority and minority charge storage capacitances must be separately computed and the storage capacitance of the trapped charge must be included in the equivalent circuit model. As indicated previously, the trapped charge is not important in almost all practical VLSI applications. Only in a few production devices (high speed switching diodes, high power rectifiers and SCR) will the stored charge on the traps be significant, but these applications require a switching model (sections 55n) instead of a small-signal model.

542 Small-Signal Numerical Examples of a Si p/n Junction

Let us compute the numerical values of these circuit elements in a Si p/n junction diode under three bias voltage conditions to illustrate their relative importance: a reverse bias ($V=-1V$), zero bias, and a forward bias ($V=+0.5V$) where minority carrier injection is important. We shall use the numerical constants employed earlier: $N_{AA}=10^{18}cm^{-3}$, $N_{DD}=10^{14}cm^{-3}$, $T=297K$, $kT=0.0255eV$, $D_n=25cm^2/V$, $D_p=9cm^2/s$, $\tau_n=\tau_p=10^{-6}s$, $\varepsilon_s(Si)\simeq 10^{-12}F/cm$. The G-V and C-V curves are plotted in Fig.542. Numerical results at three voltages are tabulated in

Table 542.1. The following conclusions can be drawn from the table where the dominant terms are in **bold**. (1) For reverse and zero biases, C is mainly from the depletion layer capacitance, C_d and G is mainly from the recombination-generation loss in the space-charge layer, G_{pn}. (2) For large forward bias, V = +0.5V, both C and G are dominated by the recombination of the injected minority carriers on the lowly doped side, in this case of a p+/n diode, C_p and G_p. This is due to the exponential dependence of C_p and G_p on the applied voltage.

Conclusions in (1) are valid also for low forward bias until those in (2) take over. The forward voltage at which solutions in (1) and solutions in (2) cross, or when G_p and C_p take over, is approximately the **injection threshold voltage** which was defined in (535.7) and estimated to be about 0.26V. This is the forward d.c. voltage where the injected minority-carrier diffusion-recombination Shockley current (holes in the n-type layer) overtakes the space-charge layer SNS recombination current.

TABLE 542.1
Computed Small-Signal Parameters
of a Si p/n Diode at Three Applied d.c. Voltages
(Dominant terms are in bold.)

	Unit	V=-1.0V	V=0v	V=+0.5V
G_n	nS/cm^2	2.92×10^{-20}	0.00314	1.03×10^{6}
G_p	nS/cm2	1.75×10^{-16}	18.8	**6.16×10^{9}**
G_{pn}	nS/cm^2	**101**	**4345**	3.73×10^{7}
C_n	pF/cm^2	2.92×10^{-23}	3.14×10^{-6}	1.03×10^{3}
C_p	pF/cm^2	1.75×10^{-19}	1.88×10^{-2}	**6.16×10^{6}**
C_{pn}	pF/cm^2	0.20	8.707	7.45×10^{4}
C_d	pF/cm^2	**2.40×10^{3}**	$3.73\times10^{+3}$	6.89×10^{3}
x_{pn}	μm	4.31	2.77	1.50
J_{n0}[exp(qV/kT)-1]	A/cm^2	-8.0×10^{-14}	0	2.62×10^{-5}
J_{p0}[exp(qV/kT)-1]	A/cm^2	-4.8×10^{-10}	0	**1.57×10^{-1}**
J_2[exp(qV/2kT)-1]	A/cm^2	**-6.90×10^{-7}**	0	4.35×10^{-3}

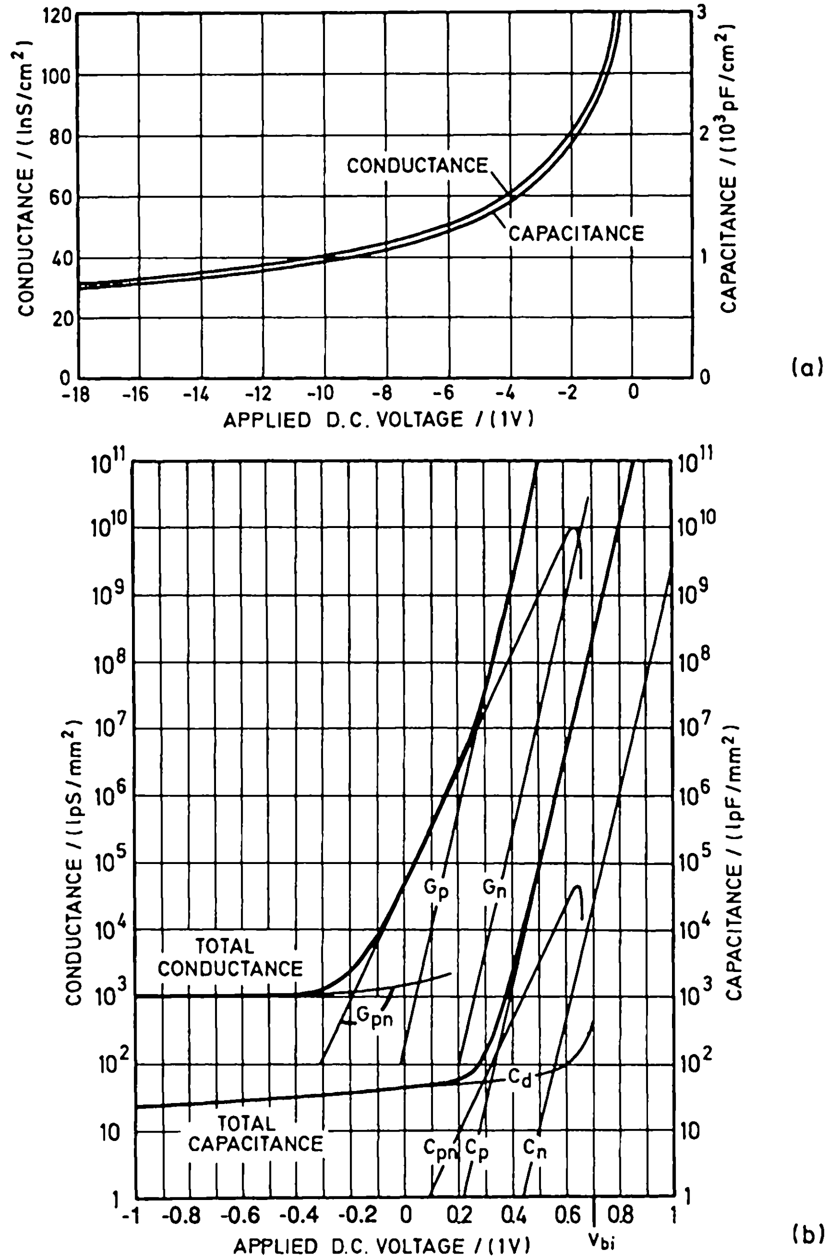


Fig.542 Calculated small-signal parameters of a p/n junction diode as a function of the applied d.c. voltage. **(a)** Reverse bias, linear C and G vs. linear V. **(b)** Small reverse bias and normal forward bias ranges, semilogarithmic or $\log_e C$ and $\log_e G$ vs linear V.

550 SWITCHING TRANSIENTS IN A P/N JUNCTION

P/N junction diodes are used in many switching circuits. The rates at which a p/n junction diode can be switched on to the high forward conduction state and off to the low reverse conduction state are obviously the most important factors which set the upper limit of the speed of a circuit. These rates are mainly controlled by the minority carrier recombination rate or lifetime in the lower doped of the two quasi-neutral layers of a p+/n or n+/p asymmetrical diode. The lower doped quasi-neutral layer is known as the **quasi-neutral base layer**. In a p+/n junction diode, the n-layer is the quasi-neutral base layer. The term 'base' comes from the **base (substrate) material** on which a p/n junction is fabricated and is analogous to a similar usage in the term, transistor base layer. In the numerical p+/n diode example given previously, it is the hole lifetime in the n-type base layer that controls the switching speed. These holes are injected into the n-type base layer from the p+ emitter layer by the forward applied voltage or current. It should be obvious that this ought to be the limiting factor because hole current in the n-type region is the dominant mechanism that controls the forward diode current at high forward currents in the thick p+/n diode. Since the p+ layer is the source of the injected holes in the n-layer, the p+ layer is known as the **emitter layer**. The layer between the front contact surface, $-T_p$, and the p-type edge of the space-charge layer, $-x_p$, is known as the **quasi-neutral emitter layer**.

When the diode is switched on, applications require that it be switched on very hard, i.e. switched to a very high conductance, so that it offers a very small series resistance to the two points in the circuit connected by the diode, essentially short-circuiting the two points. High conductance is obtained at high forward current. High forward current is obtained at a forward applied voltage near the diffusion or built-in voltage, V_{bi}, in order to diminish the potential barrier across the p/n junction space-charge layer since the barrier limits the flow of electrons and holes. This gives a simple rule-of-thumb design rule: it takes about V_{bi} or 0.7V [from (522.3)] to switch on a p/n diode. The frequently heard value from researchers and circuit designers is 0.8V because N_{DD} and N_{AA} in most practical diode switches and rectifiers and in the emitter-base junction of most bipolar transistors are 10-times higher, $N_{DD}\approx10^{19}$ and $N_{AA}\geq10^{16}\text{cm}^{-3}$, increasing V_{bi} by $(kT/q)\log_e 100=4.6kT/q\approx0.12\text{V}$, and because of the voltage drop across the series resistances.

The **switch-on** or **turn-on** transient current and voltage waveforms of a p/n junction diode are determined by how fast the minority carriers can be injected to charge up the quasi-neutral n-type base layer in a p+/n diode or p-type base layer in a n+/p diode. The switch-off or turn-off transient is determined by how fast these injected minority carriers can be extracted or diminished to the very low concentrations at zero or reverse biases. Three methods are generally used to increase the rate of reducing the concentration of the injected minority carriers. They are: (i) increasing the density of recombination centers, (ii) using a thin base layer with substrate contact of high recombination rate, known as a good **ohmic contact**, and (iii) using a thin base layer on a highly doped substrate such as the diode structure p+/n/n+ whose n-base layer is very thin. The n/n+ junction in the p+/n/n+ diode structure of (iii) is known as the **majority carrier contact**. It was also known as the **donor contact** coined by R. N. Hall in 1952. Although the thin-base diode structure is more often encountered in two-terminal diodes and integrated circuits, we will analyze only the thick-base diode because it brings out the essential physics using only simple arithmetic. Series resistances may become a speed limitation if the R_sC time constant (C is the diode capacitance) becomes large and comparable to the minority carrier lifetime or the desired switching delay time.

The ancient descriptions of the p+/n/n+ diode structures of (iii) above were 'narrow-base' and 'wide-base' diodes borrowed from the term 'transistor base width' originated by Shockley during his invention of the bipolar junction transistor theory in 1949. These term must be discarded in favor of the modern terms, **thin-base and thick-base diodes**, and the corresponding terms, **thin-emitter and thick-emitter diodes**. The old terminology will hinder intelligent communication among engineers and manufacturers who are producing transistors and integrated circuit chips today. The reason is geometrical. The modern small-dimension (submicron) diodes and transistors used in integrated circuits employ thin and thick layers whose thicknesses (0.1μm to $< 1\mu$m) are much smaller than the areal dimensions ($> 1\mu$m to 100μm) a few to a few tens micrometers). Thus, the descriptives, narrow and wide, must be used for the areal dimensions of the junction area instead of the thickness of the junction layers which must be described as thin and thick.

In addition to the large current and voltage switching transients just described, there are also very small and rather slow reverse current (or voltage) and capacitance transients in p/n junctions due to trapping in the space-charge layer, just like those in the MOSC described in sections 42n but simpler since only phase I exists. Although the reverse current-voltage and capacitance transients are small, hence unimportant in circuit applications, they have given the most powerful technique to measure the property of traps which are described in section 554.

551 Charge-Control Switching Analysis of a p/n Junction

Switching current and voltage waveforms can be calculated by solving the time-dependent continuity equations of electrons and holes given by (350.1) to (350.6). However, simple analytical solutions cannot be obtained in general unless simplifying assumptions are made. These assumptions again lead to the **charge control equation** and give the **charge control model**.

We shall analyze a simple one-dimensional case whose solutions can be obtained analytically in order to bring out the physics and illustrate the analytical technique. Consider the thick-base p+/n junction diode. The diode current is limited by diffusion and drift in the lower-doped n-base and not the higher-doped

p^+ emitter. We shall also assume that the current in the n-layer is dominated by diffusion of the minority carriers (holes), $-qD_p dP/dx$, and that the majority carrier (electron) drift current, $q\mu_n NE$, is not important because the electric field in the quasi-neutral n-layer, E, is small even though the majority carrier concentration, N, is high. We shall also use the linear recombination-generation law represented by a constant lifetime, τ_p. The last two assumptions were the very ones used to obtain the d.c. and small-signal characteristics of the Shockley diode. The hole continuity and hole current equations were given by (350.2) and (350.4) and are

$$q(\partial p/\partial t) = -\partial j_P/\partial x - q(p - P_N)/\tau_p \tag{551.1}$$

$$j_P = - qD_p(\partial p/\partial x) \tag{551.2}$$

Substituting j_P of (551.1) into (551.2), we have

$$q(\partial p/\partial t) = + qD_p(\partial^2 p/\partial x^2) - (p - P_N)/\tau_p \tag{551.3}$$

This differential equation can be readily solved using the Laplace Transform technique to give the space-time dependences of the hole concentration, p(x,t), in the quasi-neutral n-type base layer ($x=x_n$ to $x=T_n$) under a given set of initial and boundary conditions. However, the mathematics is fairly complicated and solutions will be postponed to bipolar transistor switching analyses in chapter 7.

Instead of seeking p(x,t) at all points in the quasi-neutral n-layer, we calculate the excess minority carrier charge, ($p-P_N$), stored in the entire quasi-neutral n-layer from $x=x_n$ to $x=T_n$. The excess hole (minority carrier) charge, stored in a slice of the quasi-neutral n-layer from x to the back surface $x=T_n$, is then

$$q_P(x,t) = q \int_x^{T_n} [p(x,t) - P_N]dx. \tag{551.4}$$

The total excess hole charge stored in the entire quasi-neutral n-layer is then

$$q_P(x_n,t) = q \int_{x_n}^{T_n} [p(x_n,t) - P_N]dx. \tag{551.5}$$

Integrating the continuity equation (551.1) from $x=x_n$ to T_n using (551.5), gives

$$\partial q_P/\partial t = j_P(x_n,t) - j_P(T_n,t) - q_P/\tau_p. \tag{551.6}$$

Rearranging the terms, we obtain the charge control equation which is

$$\frac{\partial q_P}{\partial t} + \frac{q_P}{\tau_p} = j_P(x_n,t) - j_P(T_n,t). \tag{551.7}$$

This is a linear differential equation. It can be solved for a given set of hole currents at the two boundaries of the quasi-neutral n-layer: $j_P(x_n,t)$ at the edge of the space-charge layer, and $j_P(T_n,t)$ at the base surface.

Although we do not know the general form of $j_P(x_n,t)$ and $j_P(T_n,t)$, since they are the unknowns we are seeking, fortunately the external circuit current is usually specified or can be measured by a meter and an oscilloscope. For the thick-base diode considered here, the contacts are nearly at equilibrium because of high generation-recombination rates, thus, the minority carrier current flowing at the contacts are nearly zero. This is true also in a thin-base diode with majority carrier contacts for a different reason. Thus, we can set $j_P(T_n,t)=0$. The external circuit current is then equal to the minority carrier current injected into the base from the p/n junction, $j_P(x_n,t)$. We shall analyze the turn-on and turn-off transients separately in the following subsections.

552 Turn-on Transient in a p/n Junction

Let a constant current, J_F(A/cm^2), be applied suddenly through the diode. What then is the time required for the junction voltage to reach its steady state value, V_F? (Why not a constant voltage step? See problems P552.n.) To answer, let the instantaneous junction voltage be denoted by $v_J(t)$. Then, using $j_P(x_n,t)=J_F$ and $j_P(T_n,t)=0$, (551.7) becomes

$$\partial q_P/\partial t + q_P/\tau_p = J_F. \qquad 0 < t < \infty \qquad (552.1)$$

Its solution is

$$q_P(t) = J_F\tau_p[1 - \exp(-t/\tau_p)]. \qquad (552.2)$$

The steady-state value of the excess hole charge stored in the n-layer is then

$$Q_P = q_P(t=\infty) = J_F\tau_p. \qquad (552.2A)$$

The forward d.c. steady-state current, J_F, is given by the hole term of the Shockley diode equation, (532.12), since in this switching example we have assumed that minority (hole) diffusion current in the quasi-neutral n-type layer is the current limiting mechanism. Thus,

$$J_F = J_{P0}[\exp(qV_F/kT) - 1] = (qP_NL_p/\tau_p)[\exp(qV_F/kT) - 1]. \qquad (552.3)$$

which is substituted into (552.2A) to give

$$Q_P = (qP_NL_p)[\exp(qV_F/kT) - 1]. \qquad (552.4)$$

This solution agrees with integrating $P(x)-P_N$ using (532.10) which gives

$$Q_P = q\int_{x_n}^{T_n}[P(x) - P_N]dx = (qP_NL_p)[\exp(qV_F/kT) - 1]. \qquad (552.4A)$$

To compute the junction voltage waveform, we assume that the hole concentration at the n-boundary of the space-charge layer, x_n, is given by the equilibrium value raised by the Boltzmann factor of the applied voltage. This boundary condition was used to obtain d.c. current-voltage characteristics. It is

$$p(x_n,t) = P_N\exp[v_J(t)/kT]. \qquad (552.5)$$

We also assume that the shape of $p(x,t_1)$ at the instance of time t_1 and at the junction voltage $v_J(t_1)$ is the same as that of the steady-state $P(x)$ calculated at the same junction voltage, $V_J=v_J(t)$. This is known as the quasi-static approximation. It assumes that the entire $p(x,t)$ profile increases with time according to (552.5) and that there is no delay in the build-up of the hole (or minority carrier) concentration with time at a position in the quasi-neutral n-layer, x, which is at a distance $x-x_n$ from the hole injection boundary x_n. The omission of the minority carrier diffusion delay from x_n to x in this approximation is the basis of all the Charge Control Models of diode and transistor theories. Thus, in this model, we have

$$q_P(t) = (qP_NL_p)\{\exp[qv_J(t)/kT] - 1\}. \tag{552.6}$$

from the steady-state result, (552.4). Equating (552.6) to (552.2), we get

$$\exp[qv_J(t)/kT] - 1 = [\exp(qV_P/kT) - 1][1 - \exp(-t/\tau_p)] \tag{552.7}$$

which can be solved for $v_J(t)$ to give

$$v_J(t) = (kT/q)\cdot\log_e\{1 + [\exp(qV_P/kT)-1][1-\exp(-t/\tau_p)]\}. \tag{552.8}$$

Generally $V_P>>kT/q$ in practical applications, so when $v_J(t)>>kT/q$ we can drop 1 relative to the exponential. This simplifies the results to:

$$\begin{aligned} v_J(t) &\approx V_P + (kT/q)\cdot\log_e[1 - \exp(-t/\tau_p)] \\ &\approx V_P - (kT/q)\exp(-t/\tau_p) \end{aligned} \tag{552.9}$$

where we used the approximation for the logarithmic: $\log_e(1+x)\approx x$ when x is small. This result shows that when the junction is switched on by a constant current, the junction voltage rises exponentially towards its steady-state value of V_P with a time constant of τ_p. This charge analysis result is a good approximation to the exact solution of the partial differential equation (551.3) which is

$$p(x_n,t) = P_N\exp(qV_P/kT)\mathrm{erf}\sqrt{t/\tau_p} = P_N\exp[qv_J(t)/kT]$$

or

$$v_J(t) = V_P + (kT/q)\log_e[\mathrm{erf}\sqrt{t/\tau_p}]. \tag{552.10}$$

Here, erf is the error function. Its complement, erfc(Z) = 1-erf(Z), was defined by (511.5A), and was encountered in atomic diffusion for fabricating diffused p/n junction described in sections 51n.

The solutions are sketched in Figs.552.1(b) and (c) where the excess hole concentration, $p(x,t)-P_N$ vs x, and the hole current density, $j_P(x,t)$ vs x, are given at four times, 0, 1, 2 and ∞. The $t=\infty$ curve is the steady-state solution. Initially the excess hole concentration and the hole current in the n-region are zero. The initial total current, J_F, is then entirely carried by electrons (the majority carriers in the n-base layer) since $j_P(x,t)+j_N(x,t)$ = constant inside the semiconductor because it must be equal to the current flowing in the external wires, J_F, according to

Kirchoff's current law. Note that the entrance slope, dp/dx at $x=x_n$, in Fig.552.1(c) is a constant at all times or in all four curves since $-D_p(dp/dx)=J_F=$constant at $x=x_n$.

The current and voltage waveforms of this turn-on transient are shown in Figs.552.2(a) and (b). Note the exponential-like rise of the junction or terminal voltage towards the steady-state value, V_F.

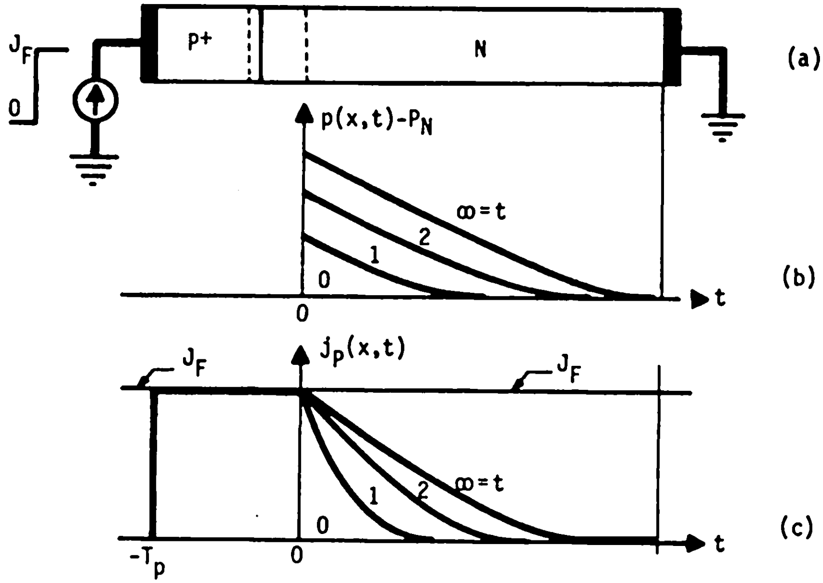


Fig.552.1 Sketches of the turn-on transients in the quasi-neutral n-layer when a p+/n junction diode current is switched from 0 to a forward value of J_F. (a) Biasing circuit. Spatial variation of the injected minority carrier (holes) (b) excess concentration and (c) current.

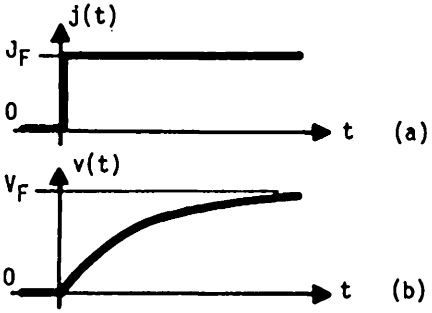


Fig.552.2 Turn-on junction voltage transient in a p+/n diode. (a) Applied current waveform. (b) Junction voltage waveform.

553 Turn-Off Transient in a p/n Junction

The charge control model can also be used to analyze the turn-off transient of a p/n junction diode. However, the waveform is much more complicated. A typical current and voltage waveform is shown in Figs.553.1(b) and (c). In this example, the diode is initially at an on state with a forward current J_F at a forward voltage V_F. At $t=0$, the diode voltage is switched to a large reverse value, V_R, which is in series with a large resistance, R. The diode current will immediately reverse itself, but because of the series resistance, it is limited to a magnitude of $J_R=V_R/R$ right after the switch as indicated in Fig.553.1(b). The reverse current is nearly constant for a period of t_I, which is known as the **constant reverse current phase** or phase I, labeled in Fig.553.1(b). During phase I, the junction voltage decreases from $+V_F$ towards $-V_R$. At the end of phase I, the junction voltage crosses zero. The period or duration of the constant reverse current phase, t_I, is defined as the time required for the junction voltage to drop to zero from V_F as illustrated in Fig.553.1(b). After t_I, the diode current will decay rapidly to $-J_1$ which is essentially zero when compared with J_F. This is known as the decaying phase or phase II, and is labeled t_{II} in Fig.553.1(b) on a later page.

The observed current waveform shown in Fig.553.1(b) can be used to guide us to make the necessary mathematical approximations in order to get good analytical solutions for t_I and t_{II}. This is an excellent example where experimental results guide the development of a theory. This was given by R. H. Kingston of M.I.T. Lincoln Laboratory in 1954 and is the earliest example of mathematical approximation based on experiments in the evolution history of semiconductor device theory. From elementary theory of differential equation, the solution of the first order differential equation of the stored charge given by (551.7) is

$$q_P(t) = \exp(-t/\tau_p)\left[\int_0^t \exp(t/\tau_p)[j_P(x_n,t) - j_P(T_n,t)]dt + K \right]. \qquad (553.1)$$

For this problem, $q_P(t=0) = Q_F = J_F\tau_p$ from (552.2) so that $K=Q_F=J_F\tau_p$. We also used $j_P(T_p,t)=0$ since there is no hole current at the distant $(T_p\to\infty)$ ohmic contact on the back surface of the quasi-neutral n-layer.

For the reverse phase, $j_P(x_n,t)=-J_R$. Thus, the integral can be evaluated explicitly, giving

$$q_P(t) = (J_F + J_R)\tau_p\exp(-t/\tau_p) - J_R\tau_p. \qquad 0 < t < t_I \qquad (553.2)$$

A charge storage time can be defined as the time required to decrease the excess stored charge to zero, assuming a constant reverse current flow at all times. The constancy of the reverse current is obviously not an attainable condition as we can see from the reverse current transient shown in Fig.553.1(b). But it is the approximation which allows us to derive a simple analytical solution from (553.2) without using complex mathematics which would obscure the technique and physics. With this definition and the constant reverse current approximation, the charge storage time, t_S, is obtained by setting $q_P(t_S)=0$. From (553.2), we get

$$t_S = \tau_p \cdot \log_e[1 + (J_P/J_R)]. \tag{553.3}$$

This charge storage time is somewhat larger than t_I since t_I does not include the decaying phase or phase II. It is, however, slightly smaller than t_I+t_{II} because of the constant reverse current assumption. The actual reverse (recombination) current decays during phase II as indicated in Fig.553.1(b), therefore, the constant current underestimates the time required to sweep out or to diminish the stored minority carrier (hole) charges in the quasi-neutral n-layer. The approximate result defined by a charge storage time, t_S, is often used in p/n diode switching circuit designs. It may be compared with the exact solutions of the partial differential equation (551.7) which are

$$t_I = -\tau_p \cdot \mathrm{erf}^{-1}[J_P/(J_P+J_R)]^2 \tag{553.4}$$

and

$$j_P(t) = -J_P[-\mathrm{erfc}\sqrt{T} + \exp(-T)/\sqrt{\pi T}]. \qquad 0< T=t/\tau_p <T_{II} \tag{553.5}$$

The phase II time constant, t_{II}, can then be obtained numerically from (553.5) if a specific definition is made for t_{II}. In Fig.553.1(b), it is defined as $j_P(t=t_{II}) = -0.1J_R$, that is, the reverse current has dropped to 10% of its initial value. This definition then gives a transcendental equation for t_{II} or $T_{II}=t_{II}/\tau_p$ which is

$$0.1J_R/J_P = -\mathrm{erfc}\sqrt{T_{II}} + \exp(-T_{II})/\sqrt{\pi T_{II}}. \tag{553.6}$$

Curves of the three solutions are given Fig.553.2. The axes are logarithmic in order to show many decades of time constants and current ratios commonly encountered in practice.

As a numerical example, consider the diode given in the previous section with a minority carrier (hole) lifetime of 1-μs in the quasi-neutral n-type base layer. Let $I_F=10$mA and the diode be switched back to $V_R=5$V in series with 50 ohms, then, $I_R=100$mA, so $I_F/I_R=0.1$. The switching delay time constants are then

$$\begin{aligned} t_I &= 10^{-6} \times 7.8754\times10^{-3} = 7.9 \text{ nano-seconds} \\ t_{II} &= 10^{-6} \times 9.0 \times10^{-2} = 90 \text{ nano-seconds} \\ t_I + t_{II} &= 7.9 + 90 = 97.9 \text{ nano-seconds} \\ t_S &= 10^{-6} \times 9.5 \times10^{-2} = 95 \text{ nano-seconds.} \end{aligned}$$

These verify the earlier conclusion, $t_I<t_S<t_I+t_{II}$, based on simple device physics. The numbers show that t_S is an excellent estimate of the turn-off delay. They also show that a large reverse current can reduce the turn-off delay substantially below the recombination lifetime, τ_{p0} or τ_{n0}: by the ratio J_P/J_R.

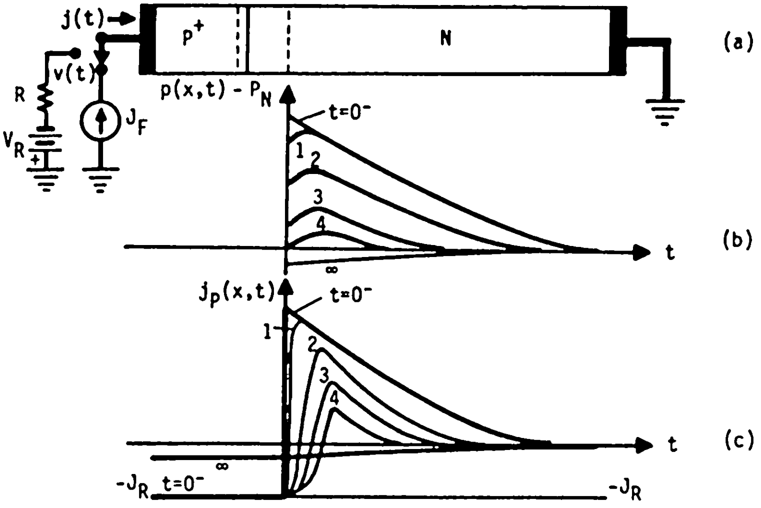

j(t)
v(t)
R
V_R
J_F
p+
N
(a)
p(x,t) - P_N
t=0⁻
1
2
3
4
∞
t
(b)
j_p(x,t)
t=0⁻
1
2
3
4
t
(c)
∞
$-J_R$ t=0⁻
$-J_R$

Fig.553.1 The turn-off (a) current and (b) voltage waveforms in a p+/n junction.

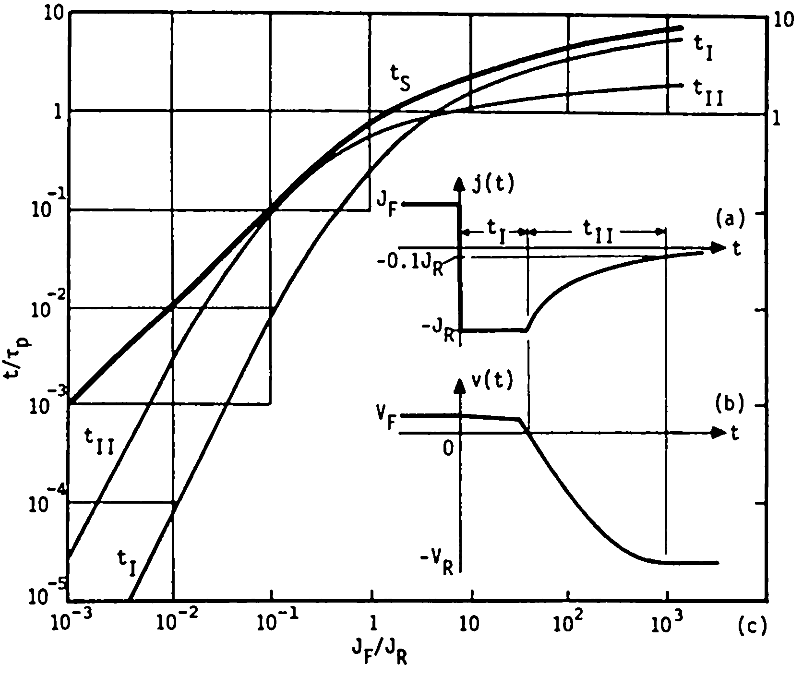


Fig.553.2 The phase I and II durations, t_I and t_{II}, of the turn-off transient in p+/n junctions.

554. Capacitance and Current Trapping Transients in a p/n Junction

The small capacitance and current trapping transients in a p/n or m/s junction come from change of the trapped charge density due to stimulated emission of the trapped electrons and holes in the space-charge layer. The stimulation can be a voltage step or a pulse of extrinsic light ($h\nu < E_G$). These dark and photo capacitance and current transients have been used to measure the properties (E_T, e_n, e_p, c_n, and c_p) of impurity and defect centers as a function of temperature and electric field. Their sensitivity is one trap in 10^4 dopant (B,P,As or Sb) so 1000 trapped electrons or holes can be detected in n+/p, p+/n or m/s diodes. The tremendous sensitivity has also made the method a very powerful diagnostic and monitoring technique in VLSI manufacturing in addition to fundamental research.

The method was proposed by this author in 1965 after recognizing the origin of these small junction transients. It was then given as homework problem on March 11, 1965 while he was teaching an undergraduate device course to junior EE students at the University of Illinois in Urbana-Champaign. The author had earlier shown (1963) that the same SRH emission-detrapping transients in the gate-junction space-charge layer was the new origin of the low-frequency $1/f^2$ and 1/f noise in junction-gate field-effect transistors [350.5]. The first experimental demonstration was the photocurrent transient observed on December 17, 1966 by his doctoral student, Al F. Tasch, in gold-doped Si p+/i/n+ diodes [554.1]. The methodology was analyzed in great detail and demonstrated experimentally during 1967-1970 [554.2]. More than twenty transients were analyzed by varying the applied voltage and photon energy. The first experimental demonstrations are reproduced in Figs.554.1(a)-(d). Instrument automation of the methods was first given in 1971 by another former doctoral student, Leopold D. Yau [554.3] which has since been automated by analog and digital computers and called DLTS. Yau's original (1971) manual capacitance transient spectrograph (CTS) from Ag-levels in Si and later digital-computer (HP-2112) automated CTS from Au-levels in Si are shown in Figs.554.2(a)-(d). The first fundamental application was the measurements of the thermal emission rates of electrons and holes at the two gold energy levels in Si shown in Fig.381.2. The extension to MOS capacitors was demonstrated by this author and his doctoral student H. S. Fu in 1972 [420.1-420.2].

[554.1] C.T.Sah and A.F.Tasch, Jr. "Precise determination of the multiphonon and photon carrier generation properties using the impurity photovoltaic effect in semiconductors," Physical Review Letters,19, 69-71, 10 July 1967.

[554.2] C.T.Sah, L.Forbes, L.L.Rosier, and A.F.Tasch,Jr. "Thermal and optical emission and capture rates and cross sections of electrons and holes at imperfection centers in semiconductors from photo and dark junction current and capacitance experiments," Solid-State Electronics,13, 759-788, June 1970.

[554.3] L.D.Yau and C.T.Sah,"Thermal ionization rates and energies of electrons and holes at silver centers in silicon," Physica Status Solidi (a)6, 561-573, 16 August 1971.

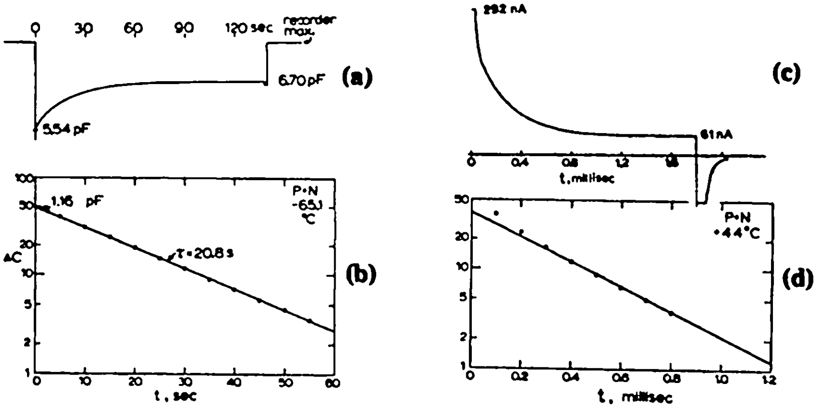


Fig.544.1 The historical first dark capacitance [(a) linear (b) semilog] and current [(c) linear (d) semilog] transients observed in Au-diffused p+/n Si diodes [544.2].

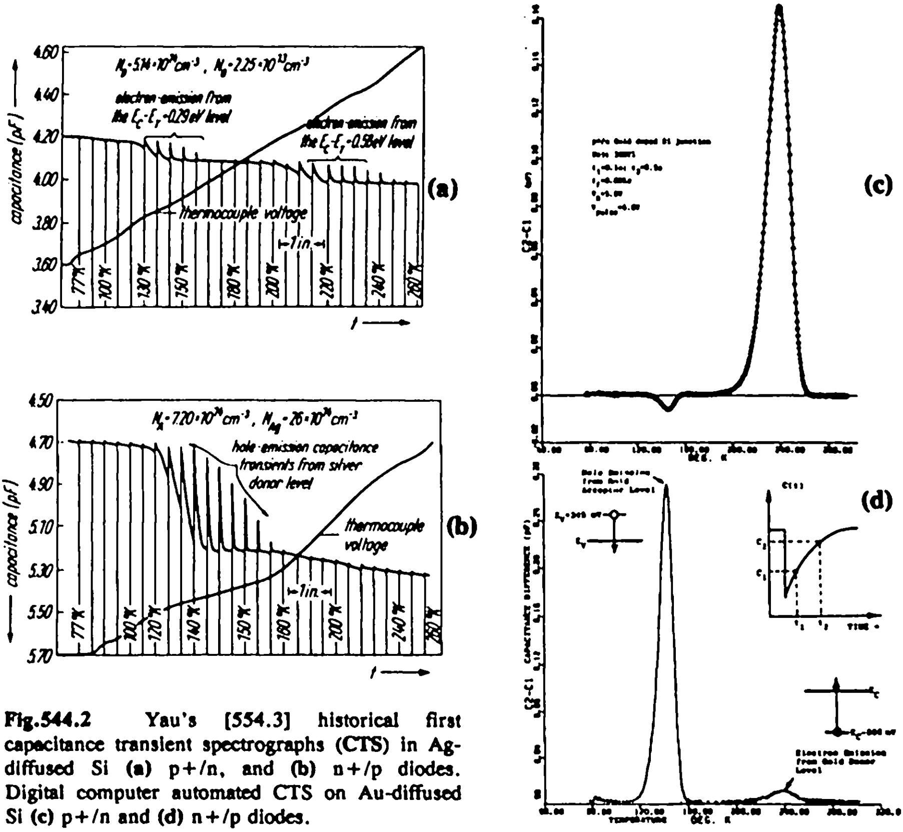


Fig.544.2 Yau's [554.3] historical first capacitance transient spectrographs (CTS) in Ag-diffused Si (a) p+/n, and (b) n+/p diodes. Digital computer automated CTS on Au-diffused Si (c) p+/n and (d) n+/p diodes.

Yau's manually operated automatic capacitance transient spectroscopy method began its popularity journey three years later in 1974 via a former Illinois postdoc, David Lang, who used an analog computer while starting a career at the Bell Telephone Laboratories. The methods have since been known as the **deep level transient spectroscopy** (DLTS), **capacitance transient spectroscopy** (CTS) and **current transient spectroscopy** (ITS). They have been used in no less than a thousand research journal articles and a hundred doctoral theses world-wide. Automated analog and digital dark capacitance transient spectrometers with 10^{-4} sensitivity have been commercially marketed since early 1980's and popularly used in basic research on deep-energy-level traps in compound semiconductors and Si.

The capacitance and current transients in a reverse biased p/n junction diode are just the phase I transients in a MOSC described in sections 42n. The inversion phase (phase II) of MOSC transient is not present in p/n junctions because there is no insulator or oxide layer to allow minority carrier accumulation and surface inversion. Thus, the phase I capacitance and current waveforms, Figs.420.1(e) and (g), and formulae, (421.8) and (421.11) for C(t), and the second term in (422.4) for i(t), are all applicable to p+/n junctions if we let $x_o \rightarrow 0$ or $C_o \rightarrow \infty$.

The formulae for the trapping transient at interface traps, (421.8) and the first term in (422.4), are also applicable to p+/n junctions if a high concentration of traps is confined in a thin layer inside the space-charge layer at a distance x_o from the p+/n metallurgical boundary. They are particularly useful for semiconductor heterojunctions since there is frequently a high density of interfacial traps near the heterojunction interface between the two semiconductors. These interfacial traps are physical defects (dangling bonds) due to lattice mismatch generated by thermal differential contraction during heterojunction formation. If the traps are located at the interface, then x_o is the thickness of the space-charge layer on the p+ side.

As a numerical example, assume that the Si p+/n diode listed in Table 542.1 is uniformly doped with Au at a concentration of $N_{TT}(\text{Au}) = 10^{12}\text{cm}^{-3}$. This diode has $N_{DD} = 10^{14}\text{cm}^{-3}$, $N_{AA} = 10^{18}\text{cm}^{-3}$, and a built-in voltage of

$$V_{bi} = (kT/q)\log_e(N_{DD}N_{AA}/n_i^2) = 0.02585\log_e(10^{14}\text{x}10^{18}/10^{20}) = 0.714\text{V}. \qquad (554.1)$$

Assume a circular diode junction with 10mil (254μm) diameter, then the area is

$$A_{pn} = \pi D^2/4 = 5.067\text{x}10^{-4}\text{cm}^2 = 5.067\text{x}10^{4}\mu\text{m}^2. \qquad (554.2)$$

Using the depletion C_d and x_d values at -1V in Table 542.1, the values at -5V are

$$C_d(-5\text{V}) = 2.4\text{x}10^3\text{x}\sqrt{[(5+0.714)/(1+0.714)]}\text{x}5.067\text{x}10^{-4}\text{pF} = 2.11\text{pF}$$

and

$$x_d(-5\text{V}) = x_{pn} = 4.31\mu\text{m}\text{x}[\sqrt{(5+0.714)}/\sqrt{(1+0.714)}] = 7.869\mu\text{m}. \qquad (554.3)$$

Let us switch the applied voltage from 0V to -5V. At -5V, the capacitance transient is due to emission of trapped electrons (majority carriers) at the midgap Au acceptor level in the space-charge layer. It can be computed from (421.11) by setting $C_o^{-2}=0$ and interchanging n and p since (421.11) was derived for a p-Si MOSC. From Fig.381.2, $e_{nb}=1000s^{-1}$ and $e_{pb}=100s^{-1}$ at 300K for the midgap Au acceptor level in Si. The total capacitance transient is then

$$\begin{aligned}\delta C_{total} &= C_I(N_{TT}/2N_{DD})[e_{nb}/(e_{pb}+e_{nb})]\\ &= 2.11\text{x}(10^{12}/2\text{x}10^{14})[1000/(100+1000)]\text{pF}\\ &= 9.59\text{x}10^{-3}\text{pF} = 9.59\text{fF}.\end{aligned} \qquad (554.4)$$

This corresponds to the detection of

$$\begin{aligned}N_{TT}x_{pn}A_{pn} &= 10^{12}\text{cm}^{-3}\text{x}7.869\text{x}10^{-4}\text{x}5.067\text{x}10^{-4}\text{cm}^2\\ &= 3.98\text{x}10^5 \text{ trapped electrons.}\end{aligned} \qquad (554.5)$$

The current transient from the second term of (422.4) is

$$\begin{aligned}\delta j_{total}A_{pn} &= q(N_{TT}x_{pn}A_{pn})\text{x}[e_{nb}(e_{nb}-e_{pb})/2(e_{nb}+e_{pb})]\\ &= 1.609\text{x}10^{-19}\text{x}3.98\text{x}10^5\text{x}[1000(1000-100)/2(1000+100)]\\ &= 2.613\text{x}10^{-11}\text{A} = 26.13\text{pA}.\end{aligned} \qquad (554.6)$$

The magnitude of the capacitance transient can be increased by increasing the area, A_{pn}, and dopant impurity concentration, N_{DD}, in the p+/n diode. It does not vary with temperature although its decay rate, $e_{nb}+e_{pb}$, does as indicated by Fig.382.2. However, the current transient is strongly temperature dependent because it is roughly proportional to the emission rate of the majority carriers (electrons in p+/n diode), e_{nb}. Thus, raising the temperature will speed up the current transient and increase its magnitude but the total trapped electron emitted [area under i(t)-t curve] does not change.

Similar numerical estimates can be made for a species of interface traps located in a thin layer in a p+/n heterojunction using (421.8) for the capacitance transient and the first term in (422.4) for the current transient. Suppose $N_{SS}=10^{11}\text{cm}^{-2}$ gold-like interfacial traps ($e_{ns}=1000s^{-s}$ and $e_{ps}=100^{-s}$) are located at the p+/n boundary, x=0, and $x_o=x_{p+}=0.1\mu$m, then

$$\begin{aligned}\delta C_{total} &= C_I(x_o/x_I^2)(N_{SS}/N_{DD})[e_{ns}/(e_{ns}+e_{ps})]\\ &= 2.11(0.1/7.869^2)10^4(10^{11}/10^{14})[1000/(1000+100)]\text{pF}\\ &= 3.098\text{x}10^{-3}\text{pF} = 3.098\text{fF}\end{aligned} \qquad (554.7)$$

and

$$\begin{aligned}\delta j_{total}A_{pn} &= q[x_{pn}/(x_{pn}+x_o)]N_{SS}A_{pn}[e_{ns}^2/(e_{ns}+e_{ps})]\\ &= 1.61\text{x}10^{-19}[7.87/(7.87+0.1)]10^{11}\text{x}5.067\text{x}10^{-4}\text{x}1000^2/1100\\ &= 7.32\text{x}10^{-9}\text{A} = 7.32\text{nA}.\end{aligned} \qquad (554.8)$$

As expected, the current transient from 10^{11}cm^{-2} interface traps is 280 times larger than the transient current from $N_{TT}x_{pn}=10^{12}\text{cm}^{-3}\text{x}7.869\text{x}10^{-4}\text{cm}=7.869\text{x}10^8\text{cm}^{-2}$ uniformly distributed bulk traps.

560 METAL/SEMICONDUCTOR DIODE

Solid state electron device traces its origin to the nonlinear current voltage characteristics discovered by F. Braun in 1874 when he placed mercury (a metal) contacts on natural crystalline, artificial crystalline, and rough sulphides. The first theory that predicted the correct direction of rectification of the metal/semiconductor diode was given by Mott in 1939. Mott assumed a semiconductor surface layer devoid of donors, acceptors and any ions so that the electric field in the surface space-charge layer is spatially constant. He found the solution for both the diffusion and drift currents of the majority carriers through the semiconductor surface space-charge layer which has been known since about 1948 as the **Mott barrier**. Schottky and Spenke extended Mott's theory by including a donor ion whose density is spatially constant through the semiconductor surface layer. This changed the constant electric field assumed by Mott to a linearly decaying electric field. This semiconductor space-charge layer under the metal is known as the **Schottky barrier**. A similar theory was also proposed by Davydov in 1939. The 1939 Mott-Schottky-Davydov diode theory has been erroneously termed the diffusion theory by subsequent researchers and authors. Although it gives the correct direction of rectification, it has also been proven that the Mott theory and its Schottky-Davydov extension gives the **wrong** current limiting mechanism and **wrong** current-voltage formulae in silicon metal/semiconductor diode rectifiers. The correct theory was developed by Hans Bethe and reported by him in an M.I.T. Radiation Laboratory Report dated November 23, 1942. In Bethe's theory, the current is limited by **thermionic emission** of electrons over the metal/semiconductor potential barrier. Thus, the appropriate name for the metal/semiconductor diode should be **the Bethe diode** instead of the Schottky diode, since the Schottky theory does not predict the modern M/S diode characteristics correctly. The terms, Schottky barrier and Schottky diode, has been used in the last fifty years (1939-1989). It is difficult and perhaps impossible to correct the erroneous traditional terminology. But we must recognize that the correct theory is due to Bethe and not to Schottky. Thus, we shall use the following compromises: (1) the **Mott barrier** ($N_{DD}=N_{AA}=0$) and the **Schottky barrier** ($N_{DD}\neq 0$ or $N_{AA}\neq 0$) for the two assumed surface space-charge layers; (2) the **Mott diode**, the **Mott-Schottky** or the **Mott-Schottky-Davydov diode**, and the **Bethe diode** for the two diode current mechanisms; and (3) the **metal/semiconductor (m/s or M/S) diode** as the generic name for all two-terminal diodes made of a metal in contact with a semiconductor. We prefer the named terms: (i) **Bethe diode** based on the majority-carrier thermionic theory, and (ii) **Mott diode** based on the majority-carrier diffusion-drift theory where we use only Mott since he was the first to advance the theory while Schottky and Davydov only extended Mott's theory and since we do want to use Schottky to designate the Schottky surface potential barrier. We use the named terms in order to contrast with the Shockley diode based on the minority carrier diffusion theory. We shall occasionally use the term **Mott-Schottky-Davydov-Bethe** or **MSDB diode**, but more frequently, **M/S diode**, in the descriptions of the metal/semiconductor diodes. The double subscript, ms, from

Metal/Semiconductor diode, is used to denote the characteristics and parameters associated with the transition or space-charge layer of the M/S diode. This is consistent with the double subscript notation, PN and pn such as in x_{pn}, which we used to denote the parameters associated with the p/n junction diode.

The reason for the confusion in the literature arises from the fact that the Mott-Schottky-Davydov diode theory gives exactly the same kind of exponential voltage dependence of the diode current as the correct theory of Bethe. The Bethe and Mott-Schottky-Davydov diode theory both predict the highly nonlinear exponential current-voltage characteristics similar to that just derived for the Shockley p/n junction diode, $I=I_1\cdot[\exp(qV/kT) - 1]$. Under forward bias, $V>>kT/q$, the diode will pass high currents and under reverse bias, $V<<-kT/q$, it passes a negligibly small and voltage-independent constant leakage current. However, the forward and reverse current densities of the Bethe and Mott diodes of a Schottky barrier are several orders of magnitude larger than the sum of the Shockley and SNS diodes in a comparable silicon p+/n diode.

The M/S diode is composed of a metal in contact with a semiconductor. It is one of the most heterogeneous heterojunctions. A perfect M/S diode would have a crystalline interface or interfacial contact surface. However, the interface or contact surface of practical M/S diodes is atomically imperfect and porous, causing electrical leaky paths known as 'patchy', because many semiconductor bonds are dangling or not tied up by the metal atoms. In addition, there are many foreign atoms at the interface, and frequently a metal/semiconductor contact is interlaced with an interfacial layer of residual foreign atoms of one or more monolayers thick. (One monolayer has about $5\text{x}10^{14}$ atom/cm^2 or roughly 5A interatomic spacing.) A monolayer or sub-monolayer of metal oxide or semiconductor oxide is almost unavoidable unless the M/S diode is fabricated in ultra-high vacuum and the fabrication includes a surface pre-cleaning by ion bombardment to get rid of surface impurities followed by a high-temperature annealing of the bombardment damage.

The rectification property of the metal/semiconductor contacts was discovered by F. Braun in 1874 on copper and iron sulphide semiconductors with mercury metal contacts. G. W. Pickard received a patent in 1906 on a point-contact rectifier using silicon. In 1907, Pierce published a paper in Physical Review showing rectification properties of diodes made by sputtering many metals on many semiconductors. The earliest M/S diodes in electronics application occurred in the early 1920's when the cat's whisker rectifiers were used in broadcast receivers. They consisted of a pointed tungsten wire (in the shape of a cat's whisker) whose tip or point is pressed against the surface of a lead sulphide crystal. The first large-area rectifier appeared around 1926 which consisted of a copper-oxide semiconductor thermally grown on a copper substrate. Subsequently, semiconductor films, such as selenium, were evaporated onto large metal substrates to form the rectifying M/S diodes. These large area diodes were used to convert a.c. current to d.c. current in electrical power applications. During 1925-1940, M/S diodes, consisting of a pointed tungsten metal wire in contact with a silicon crystal base, were fabricated in laboratories to detect microwaves in the UHF range. A World War II program to manufacture high-purity silicon as the crystal base for the point-contact rectifier was suggested by Fredrik Seitz in 1942 and successfully undertaken by the Experimental Station of the E. I. du Pont de Nemours and Company. Doping of the purified silicon by boron in 1943 was found to produce high conductivity silicon crystals which resulted in M/S frequency converters or mixer diodes of improved sensitivity and uniformity with high resistance to burnout. These developments of manufacturing techniques made it possible to produce in large volumes of high-quality crystal rectifiers for use in the 3-30 GHz range for radars during World War II.

The use of the M/S diode rectifier was even proposed by Lilienfeld in 1926 in the first of his three transistor patents as the gate of the field-effect transistors. In the 1926 patent, the metal/semiconductor rectifying barrier was used as the gate electrode to control the conductance and current flowing through a semiconductor film. This is the modern metal-semiconductor-field-effect transistor (MESFET). Although Lilienfeld might not have understood the complete basic physics of the metal/semiconductor rectifying barrier and the field-effect transistor structures he invented, the many circuits in his patents certainly described the correct electrical operation of the devices, including the correct polarity of the applied d.c. bias and location of the input and output signals. The correct theory of the field-effect transistor using a metal/semiconductor gate was advanced by Shockley in 1939.

A metal with a higher electron work function than the semiconductor must be used to form a potential barrier at the metal/semiconductor interface. Such a potential barrier is necessary to restrict the majority carriers (electrons in n-Si) of the semiconductor from entering the metal. The potential barrier also restricts the metal electrons from moving to the semiconductor. Without such a potential barrier, the rectification characteristics will be lost and we would have an ohmic contact which conducts easily in both directions of current flow. The latter is the basis of a good ohmic contact to give low contact resistance. Low resistance contact is essential for connecting a high performance diode, transistor or integrated circuit to other devices and electronic circuits in a system.

The Bethe diode current is dominated by those majority carriers (electrons in n-type semiconductor) that have enough kinetic energy to go over the potential barrier at the metal-semiconductor interface. All modern metal/Si Schottky barrier diodes operate in the Bethe (thermionic) diode mode instead of the Mott (diffusion-drift) diode mode. The words 'go over', instead of 'negotiate', distinguishes the two electron current paths: going over the potential barrier and tunneling **through** the potential barrier. The current from electrons going over the potential barrier at thermal equilibrium is known as the **thermionic current** in analogy to the electrons emitted from the heated cathode surface in vacuum tubes such as the color or monochrome computer monitor and TV display tube, and the magnetron in a microwave oven.

Because of the potential barrier, the minority carrier injection current (holes injected from the metal into the n-Si) is so much smaller than the majority carrier thermionic current. Thus, negligible minority carriers (holes) are stored in the quasi-neutral n-Si layer. Consequently, the metal/semiconductor diode can be switched on and off at extremely high speed without the delay in removing or supplying the minority carriers because the dominant current, the majority carrier current, responds at the dielectric relaxation speed, 10^{-12} seconds in semiconductor and still faster in the metal layer. The slow rise and decay of the minority carrier charge stored in the semiconductor of the Shockley diode, computed in sections 55n, can be neglected because the minority carrier current is minute compared with the majority carrier current. The switching speed of the **Bethe diode** is mainly limited by the RC time constant of the majority carriers which charges and discharges the Schottky barrier junction capacitance, C, through the series bulk resistance of the semiconductor, R. This is very fast. The switching speed of the **Mott diode** is slower since it is limited by the additional diffusion, drift and recombination delays in the semiconductor surface-space-charge and quasi-neutral base layers. Thus, correct name (in **bold** above) must be used even more so.

The M/S diodes have been used in many recent integrated circuit applications to take advantage of their high switching speed. The Al/n-Si M/S Bethe diode has been used to clamp the collector-base junction voltage of the bipolar junction transistor so that the transistor would not be driven into the saturation range of its characteristics. This has been known as the Schottky-diode-clamped (more correctly the Bethe-diode-clamped) bipolar junction transistor. Without the M/S diode clamp, the bipolar junction transistor could be driven into saturation by a large base-input current. The collector-base junction would then be forward biased and heavily injecting minority carriers into the collector region. These minority carriers would have to be removed in order to switch off the transistor which would produce a larger delay and reduce the switching speed. This application began in 1960 and was responsible for the wide popularity and success of the low-power Schottky-barrier Bethe-diode clamped TTL (transistor-transistor logic) circuits. The use of the Schottky-barrier Bethe diode to speed up the turn-off transient in Si bipolar junction transistors has a distinct fabrication-manufacturing advantage over the technology of diffusing a recombination impurity (such as gold) into the transistor in order to kill the minority carrier recombination lifetime. This is because the gold recombination impurity would also reduce the transistor current gain, and could not be localized to the base-collector junction area of each bipolar transistor on a chip since gold diffuses rapidly in Si at high temperatures causing it to cover the entire Si wafer quickly (See Fig.512.3). After thirty years, the Schottky-barrier Bethe-diode-clamped TTL bipolar Si circuits, first introduced in 1960, are still popular in small-scale and medium-scale silicon integrated circuits. However, reproducibility and uniformity of the metal/semiconductor junctions are difficult to attain over large

Si areas due to the extreme dissimilarity between a metal and a semiconductor. The dissimilarity or heterogeneity causes interfacial diffusion of the metal and semiconductor atoms into each other's surface layer. Thus, the more easily produced MOS and CMOS integrated circuits have replaced the Schottky-clamped low-power TTL circuits in medium and high speed switching applications since 1980. In the very high speed logic and buffer memory sections of mainframe and super computers, the emitter-coupled-logic (ECL) bipolar-junction-transistor integrated circuits are used which are not driven into saturation and do not employ the Schottky-barrier Bethe-diode clamp. The importance of the Schottky-barrier Bethe diode behavior resurfaced recently in the polycrystalline emitter layer of the lastest ultra-high-speed (sub-10 picosecond) ECL bipolar junction transistors.

Silicon M/S or Bethe-Mott Schottky-barrier diodes have also been used since about 1980 in large area and high current rectifiers because the high switching speed and the negligible minority carrier storage can give very good rectification or power conversion efficiency. Reproducibility and uniformity of the M/S contact have restricted its application mainly to single diode devices and prevented its use in large-scale integrated circuits.

Another emerging application of the metal/semiconductor junction is its use as the gate junction of a junction-gate field-effect transistor on compound semiconductors. This is known as the MESFET (Metal Semiconductor Field Effect Transistor). The principal feature is that the metal gate length (the symbol L used in the MOSFET analysis) can be made extremely short (less than 0.1 micron or 1000A) by ingenious shadow evaporation and lithography techniques, resulting in single-device GaAs MESFETs with cutoff frequencies greater than 100 GHz ($1\text{GHz} = 10^9\text{Hz}$) and switching times less than 10 picoseconds. However, due to heterogeneity, large-scale integrated circuits using GaAs MESFET is still an anticipation (in 1991) after decades of engineering efforts.

561 Equilibrium Energy Band Diagrams of the Schottky Barrier

The electron energy band diagram of a metal/semiconductor Schottky barrier junction, at **electrical equilibrium** or **zero d.c. current** and **electrical-thermal-optical equilibrium** or zero voltage, is described first since it is needed to derive the dc, small-signal, and transient characteristics. The step-by-step constructions of the energy band diagram are shown in Figs.561.1(a)-(c) and to be described in detail in this section.

The meaning of equilibrium is reviewed again since it is a common practice to use terms such as thermal or thermodynamic equilibrium, or just equilibrium to denote the zero-current and zero-voltage condition. In these usages, it is tacitly understood that the metal/semiconductor structure is not strictly at thermodynamic equilibrium because of the tremendous atomic heterogeneity, from the noncrystalline metal electrode to the crystalline semiconductor. Instead, the structure is at a **quasi-thermodynamic equilibrium** or more explicitly, a nearly **structural or chemical steady state.** In this state, the atomic composition or positions of the atomic cores of the metal and semiconductor are fixed during an observation or measurement time period that is small compared with the time required to homogenize the inhomogenous distribution of the atoms by interdiffusion of the atomic cores. Except in the analysis of device reliability or aging and failure mechanisms due to structure changes from atomic migration or interdiffusion, equilibrium or electrical equilibrium means electrical and thermal-optical equilibrium or zero terminal current and zero terminal voltage.

The steps to construct the equilibrium energy band diagram are now described. Figure 561.1(a) shows the equilibrium energy band diagrams of the isolated metal on the left and isolated n-type semiconductor on the right before making a contact. The vacuum level, VL, is used as the reference energy level and set at 0 eV. This choice is suggested by the 1/r Coulomb potential energy of an electron arising from a positive nuclear point charge which approaches 0 eV when the electron is at a large distance from the positive nuclear point charge or $r\rightarrow\infty$. Thus, all the electron energies and energy levels (positive in the upward or y direction) are measured from the vacuum level. In detailed diode-transistor analyses to follow, different references are used for the quasi-Fermi potentials and the electric potential so that the analytical solutions are simple and the physics clear at first glance.

The energy levels and energies are defined as follows. F_M and F_N are the equilibrium Fermi energy levels of the electrons in the metal and n-type semiconductor respectively. W_M is the **electron work function** in the metal, frequently abbreviated as the **metal work function.** ($W_M\equiv q\phi_M$ where ϕ_M is the magnitude of the electron work function in unit of Volt.) The work function is defined as the work one must do in order to remove one electron from the material to infinity. Figure 561.1(a) shows that the electron work function in the metal is equal to the 'depth' of the Fermi energy level measured from the vacuum level since in a metal, all the energy levels below the Fermi level are filled by the available valence electrons of the neutral metal. Thus, the metal work function is also the ionization potential of the neutral metal in analogy to the ionization potential of an isolated neutral atom. The ionization potential of a material is the minimum energy required to remove one electron from the neutral metal. It has been measured using the photoelectric effect rather than by directly measuring an excess or deficiency of charge on a piece of material. Excess charge measurement is much more difficult since the excess charge will disappear in a few dielectric relaxation times which is a few picoseconds in semiconductors and orders magnitude smaller in metals.

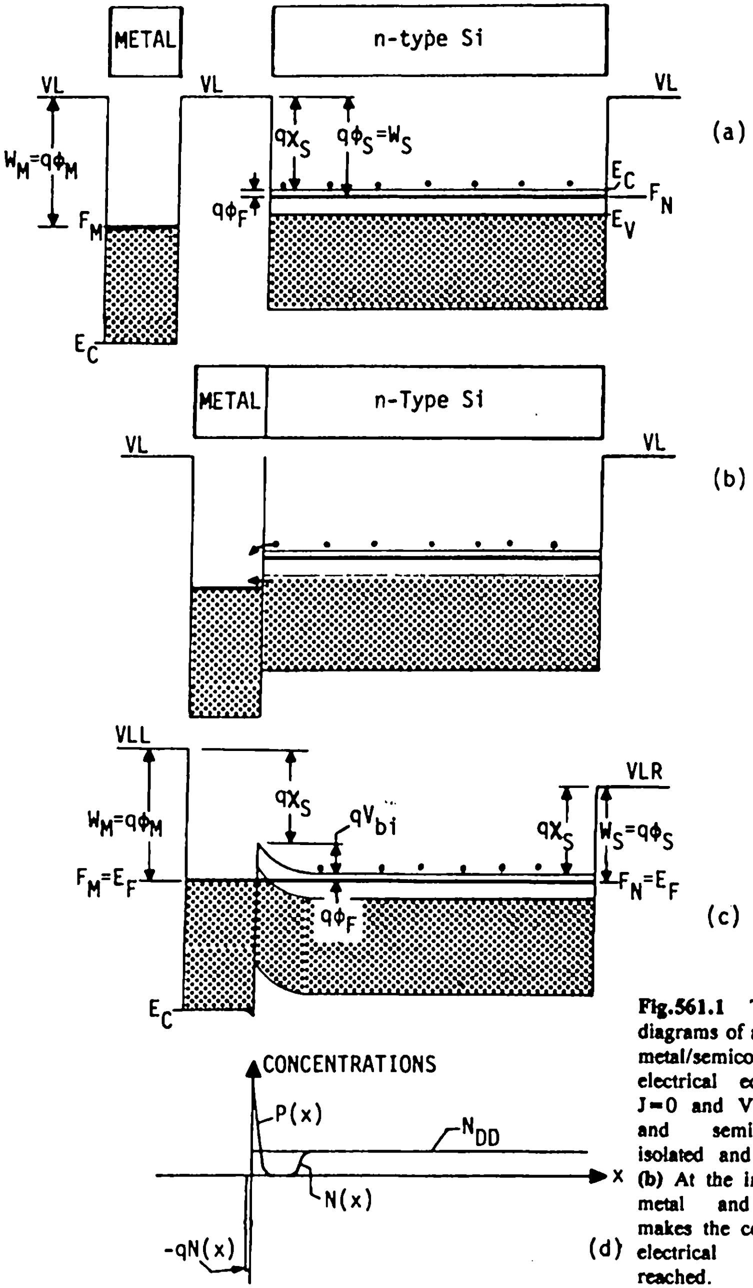

METAL
n-type Si
VL
VL
VL
$W_M = q\phi_M$
F_M
E_C
$q\chi_S$
$q\phi_S = W_S$
$q\phi_F$
E_C
F_N
E_V
(a)
METAL
n-Type Si
VL
VL
(b)
VLL
$W_M = q\phi_M$
$F_M = E_F$
E_C
$q\chi_S$
qV_{bi}
$q\phi_F$
VLR
$q\chi_S$
$W_S = q\phi_S$
$F_N = E_F$
(c)
CONCENTRATIONS
P(x)
N_{DD}
N(x)
x
-qN(x)
(d)

Fig.561.1 The energy band diagrams of a Schottky barrier metal/semiconductor diode at electrical equilibrium, i.e., $J=0$ and $V=0$. **(a)** Metal and semiconductor are isolated and not in contact. **(b)** At the instance when the metal and semiconductor makes the contact. **(c)** After electrical equilibrium is reached. **(d)** Equilibrium space-charge distribution.

Strictly, the work function of a material contains two terms, $W_M = (VL-E_C) + (E_C-F_M)$. The first term is the lower edge of the conduction band, E_C, measured from the vacuum level, VL. The second term is the metal Fermi level, F_M, measured from the lower conduction band edge. In metals, the Fermi level, F_M, is above the bottom of the conduction band edge, E_C, or $E_C-F_M<0$ so we write W_M for metal in the second and more transparent form, $W_M=(V_L-E_C)-(F_M-E_C)=(VL-E_C)$. The clearest way to show the definition of the metal work function is to label it on the energy band diagram of the metal near the metal surface. This is done in the metal energy band diagram on the left side of Fig.561.1(a).

The electron work function in the semiconductor, abbreviated **semiconductor work function**, is similarly defined. It is the depth of the Fermi level measured from the vacuum level, $W_S = q\phi_S = VL-F_N$. It also contains two terms, as shown in Fig.561.1(a): (i) the lower semiconductor conduction band edge measured from the vacuum level, $VL-E_C$, known as the electron affinity of the semiconductor whose atomic origin is illustrated in section 172 and Fig.172.3, and (ii) the electron Fermi level measured from the conduction band edge, E_C-F_N, which is positive for a nondegenerate semiconductor but can be made negative (i.e. the Fermi level is in the conduction band) if the semiconductor is heavily doped with a donor impurity. Thus, the semiconductor work function is given by $W_S = q\phi_S = (VL-E_C) + (E_C-F_N) \equiv q\chi_S + q\phi_P$. Again, the clearest illustration is a labeled energy band diagram, such as that for the n-Si given on the right side of Fig.561.1(a).

In the first step, we bring the isolated metal and semiconductor together to make electrical and atomic contact with each other. The energy band diagram at the instance of contact is shown in Fig.561.1(b). It is obvious that some conduction and valence band electrons in the semiconductor are occupying higher energy levels than the empty levels in the metal. Thus, these semiconductor electrons must move to the lower energy levels in the metal in order to reach the lowest possible energy of the metal/semiconductor system. The two left-pointing arrows in this figure show the transfer of the semiconductor electrons to the metal. These electron transfers produce an internal current and charge up the metal negatively and the semiconductor positively. The two space charges produce an internal electric field across the metal/semiconductor interfacial layer, known as the **dipole** layer or interfacial dipole layer, and a difference in potential energy or electric potential between the semiconductor and the metal, known as the **contact potential difference**. The contact potential difference is precisely also the diffusion potential or the built-in potential of a junction, such as that described for p/n junctions earlier. The charge transfer is completed (or has reached a dynamic or thermodynamic equilibrium, or steady state) when the electron drift current from the negatively charged metal to the positively charged semiconductor owing to the interfacial electric field is increased to a value which just balances the current due to electron transfer from the semiconductor to the metal to give a net zero current. This electrical equilibrium is approached quickly, in about a dielectric relaxation time if the contact interface is atomically perfect or atomically intimate, that is, without interfacial oxides and voids. When the net electron current (transfer current minus drift current) drops to zero, the metal/semiconductor contact is said to have reached **electrical equilibrium**.

The energy band diagram after the establishment of electrical equilibrium is shown in Fig.561.1(c). Notice four important features described as follows.

(1) There is now a vacuum level on the left, VLL, and a vacuum level on the right, VLR, and they differ by the work function difference between the metal and the semiconductor,

$$V_M - V_S = q(\phi_M - \phi_S) = qV_D \equiv qV_{bi} = q(VLR - VLL) \qquad (561.1)$$

or

$$V_D \equiv V_{bi} = \phi_M - \phi_S \equiv \phi_M - (\chi_S + \phi_F) = VLR - VLL. \qquad (561.1A)$$

Some authors have drawn a smooth 'vacuum' potential energy level curve, VL(x), throughout the semiconductor and the metal to join the two vacuum levels, VLL and VLR. We avoid this since it is misleading and confusing because the term 'vacuum level' is a meaningless quantity inside the metal and semiconductor and because there are two different vacuum levels separated by the metal/semiconductor sandwich.

(2) There is space-charge layer in the semiconductor side of the metal/semiconductor interface. The electron potential energy barrier height across this **semiconductor space-charge layer** is

$$qV_D = qV_{bi} = q(\phi_M - \phi_S) = q[\phi_M - (\chi_S + \phi_F)]$$

which is precisely the work function difference given by (561.1). The space-charge layer thickness can be computed using the depletion model of the p/n junction given by (523.19) and is

$$x_{mn0} = \sqrt{2\varepsilon_s V_{bi}/qN_{DD}}. \qquad (561.2)$$

The subscripts m and n denote metal and n-Si respectively and the subscript 0 denotes zero current or applied voltage. This potential barrier is created because the electrons transferred from the semiconductor to the metal had charged up the metal negatively and the semiconductor positively. Note that spatial variations of the three band edges, $E_C(x)$, $E_V(x)$ and $E_{V'}(x)$, are identical and they are parallel. This semiconductor space-charge layer is known as the **Schottky barrier** when it contains donor or acceptor ions and the **Mott barrier** when it is devoid of any ions or the semiconductor is pure or not doped with impurities.

(3) There is also a spatial variation of the potential energy of the electrons in the metal. It occurs in a very thin layer near the metal/semiconductor interface and

is depicted by the lower edge of the conduction band of the metal, $E_C(x)$, shown in Fig.561.1(c). The height and the thickness of this **metal space-charge layer** are very small due to the very high electron concentration in the metal. The thickness is roughly the Debye length which is about $25\text{x}10^{-4}\text{x}\surd(10^{10}/10^{22})=0.25\text{A}$. It is much less than the inter-electronic and interatomic spacing in the metal, $\approx(10^{22})^{-1/3}\approx4.6\text{A}$. So the energy band bending in the metal is macroscopically insignificant and is generally not shown in the energy band diagram. However, a thin layer of space-charge, a Dirac delta function or impulse function, is always drawn on the metal side of the M/S interface in the space-charge distribution diagram such as that shown in Fig.561.1(d). The area-under-the-curve of this thin metal space-charge density is exactly equal to the area of the space-charge density curve on the semiconductor side. This equality can be proved from the charge neutrality condition or Gauss theorem. But it is also the direct physical result from the basic charge transfer processes that gave Fig.561.1(c) from Fig.561.1(b) which we just described.

(4) The electron quasi Fermi level in the metal, F_M, and semiconductor, F_N, are lined up horizontally. They are labeled by one symbol, the Fermi level E_F. This is the result of electrical equilibrium or the zero current condition, $J_N=0$, which was shown in section 330 to give a continuous and horizontal (or spatially constant) Fermi level even if there is a discontinuity or step in the electron energy levels. This constancy was demonstrated mathematically by (330.1) and (331.1) which gave

$$J_N = \mu_n N(dF_N/dx) = \mu_n N(dE_F/dx) = 0. \qquad (561.3)$$

Thus, $dE_F/dx=0$ since $N\neq0$ in general. Then, E_F must be independent of position or time, i.e., it is contant in space-time.

562 DC Current-Voltage Characteristics of the M/S Diode - Bethe Theory

The electrical equilibrium energy band diagram given in Fig.561.1(c) is simplified in Fig.562.1(a) to facilitate the derivation of the current-voltage characteristics. Two references potentials are used. (i) The applied voltage is measured on the metal relative to the semiconductor using the quasi-Fermi potential of the semiconductor far from the metal/semiconductor interface as the reference, i.e., $V_N(x\to\infty)=0$. Thus, $V_M - V_N(\infty) = V_{APPLIED} = V_A$, where $V_M=F_M/(-q)$ and $V_N(x)=F_N(x)/(-q)$. (ii) The electrostatic potential, $V_I(x)=E_I/(-q)$, is measured from the far right side of the semiconductor, $V_I(x\to\infty)=0$. These are selected to simplify the mathematical derivation since to use a common reference zero for both the quasi-Fermi potentials and the electrostatic potential would drag along a constant factor in either the quasi-Fermi potentials or the electrostatic potential, making some of the expressions unwieldy and their physics harder to understand. Separate references can be selected without losing generality of the results since potential energy and electric potential are not defined absolutely and must be measured from their designated references. Thus, one can always choose arbitrary references to simplify the notation and solution of a given problem.

The energy band diagram at a forward applied d.c. voltage V_A is shown in Fig.562.1(b) and at a reverse applied d.c. voltage $-V_R$, in Fig.562.1(c). The semiconductor barrier heights at zero, forward and reverse biases are

$$V_B = V_{bi} = \phi_M - (\chi_S+\phi_P) \quad \text{when } V_A = 0 \tag{562.1A}$$

$$V_B = V_{bi} - V = \phi_M - (\chi_S+\phi_P) - V_A \quad \text{when } V_A = +V > 0 \tag{562.1B}$$

and

$$V_B = V_{bi} + V_R = \phi_M - (\chi_S+\phi_P) + V_R \quad \text{when } V_A = -V_R < 0. \tag{562.1C}$$

The d.c. current of the M/S diode can be obtained by calculating the net rate of electrons passing through the metal-semiconductor interface at $x=0$. Only those with a velocity $-v_x$ and kinetic energy greater than qV_B can pass from the semiconductor through the interface into the metal. Thus, the current component from these electrons is

$$\vec{J}_{Nx} = (-q)\int(-v_x)dn = q\int(\partial E/\hbar\partial k_x)dn \tag{562.2}$$

where $v_x=(\partial E/\hbar\partial k_x)$ is the group velocity introduced in section 190 and illustrated in Fig.190.1(b). dn is the electron density given by

$$dn = f(E)2dk_xdk_ydk_z/(2\pi)^3. \tag{562.3}$$

f(E) is the Fermi-Dirac or Fermi distribution function. It can be approximated by the Boltzmann distribution since in this example we assume that the semiconductor is not too highly doped so that the semiconductor Fermi level is several kT's below the conduction band edge, E_C, as illustrated in Figs.562.1(a),(b) and (c). Thus,

$$f(E) = 1/\{1 + \exp[(E-E_F)/kT]\} \simeq \exp\{-(E-E_F)/kT\}. \tag{562.4}$$

We also assume a simple spherical energy surface for the conduction band given by

$$\begin{aligned} E &= E_C + (\hbar^2k_x^2/2m_x) + (\hbar^2k_y^2/2m_y) + (\hbar^2k_z^2/2m_z) \\ &= E_C + (\hbar^2k^2/2m^*) \end{aligned} \tag{562.5}$$

where $m_x=m_y=m_z=m^*$ and $k^2=k_x^2+k_y^2+k_z^2$. Thus, the velocity in (562.2), $v_x=(\partial E/\hbar\partial k_x)$, simplifies to $v_x=\hbar k_x/m_x=\hbar k_x/m^*$. The integration over the two transverse components of the velocity or wave vector, k_y, and k_z, are to be extended from $-\infty$ to $+\infty$ to include all the electrons. The integration over the longitudinal component, k_x, is limited by the potential barrier since the electron kinetic energy in the x-direction, $\hbar^2 k_x^2/2m_x$, must be greater than the barrier height, qV_B. From assumed $m_x=m_y=m_z=m^*$, then the lower limit of the k_x integration is given by

$$\hbar^2 k_{x0}^2/2m^* = qV_B. \tag{562.6}$$

Substituting (562.4), (562.5) and (562.6) into (562.2), and using the definition of the definite error integral

$$\int_{-\infty}^{\infty} \exp(-x^2)dx = \sqrt{\pi}, \tag{562.6A}$$

the current density due to electrons flowing from the semiconductor to the metal (in the negative x direction) is

$$\vec{J}_{Nx} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{-k_{x0}} \exp[-(E_C-E_F)/kT]\cdot\exp[-\hbar^2(k_x^2+k_y^2+k_z^2)/2m^*kT] \cdot q(\hbar k_x/m^*)2dk_x dk_y dk_z/(2\pi)^3 \tag{562.7}$$

$$= [qm^*(kT)^2/2\pi^2\hbar^3]\cdot\exp[-q(\phi_F+V_B)/kT]$$

$$= J_{TH}\cdot\exp[-q(\phi_F+V_B)/kT] \tag{562.7A}$$

$$= J_{TH}\cdot\exp[-q(\phi_M-\chi_S-V)/kT]. \tag{562.7B}$$

To obtain (562.7B) from (562.7A), the Schottky barrier height, V_B, given by (562.6) and (562.1B) are used.

The pre-exponential factor, J_{TH} in (562.7), is known as the **thermionic current coefficient**. It is defined by

$$J_{TH} = qm^*(kT)^2/2\pi^2\hbar^3 = 120(m^*/m)T^2 \ \ \mathrm{A/cm^2} \tag{562.8}$$

$$= 10.8(m^*/m)(T/300)^2 \ \ \mathrm{MA/cm^2}. \tag{562.8A}$$

The unit MA or million amperes is used in (562.8A) since it gives an easily remembered numerical result of 11 MA/cm^2 for J_{TH} when $m^*=m$ and $T=300$K. J_{TH} given above has exactly the same form as the current due to thermionic emission of electrons from a hot cathode in vacuum. This is expected since they have identical physical origin. Thus, the current is the solid-state analog of the vacuum **thermionic current**.

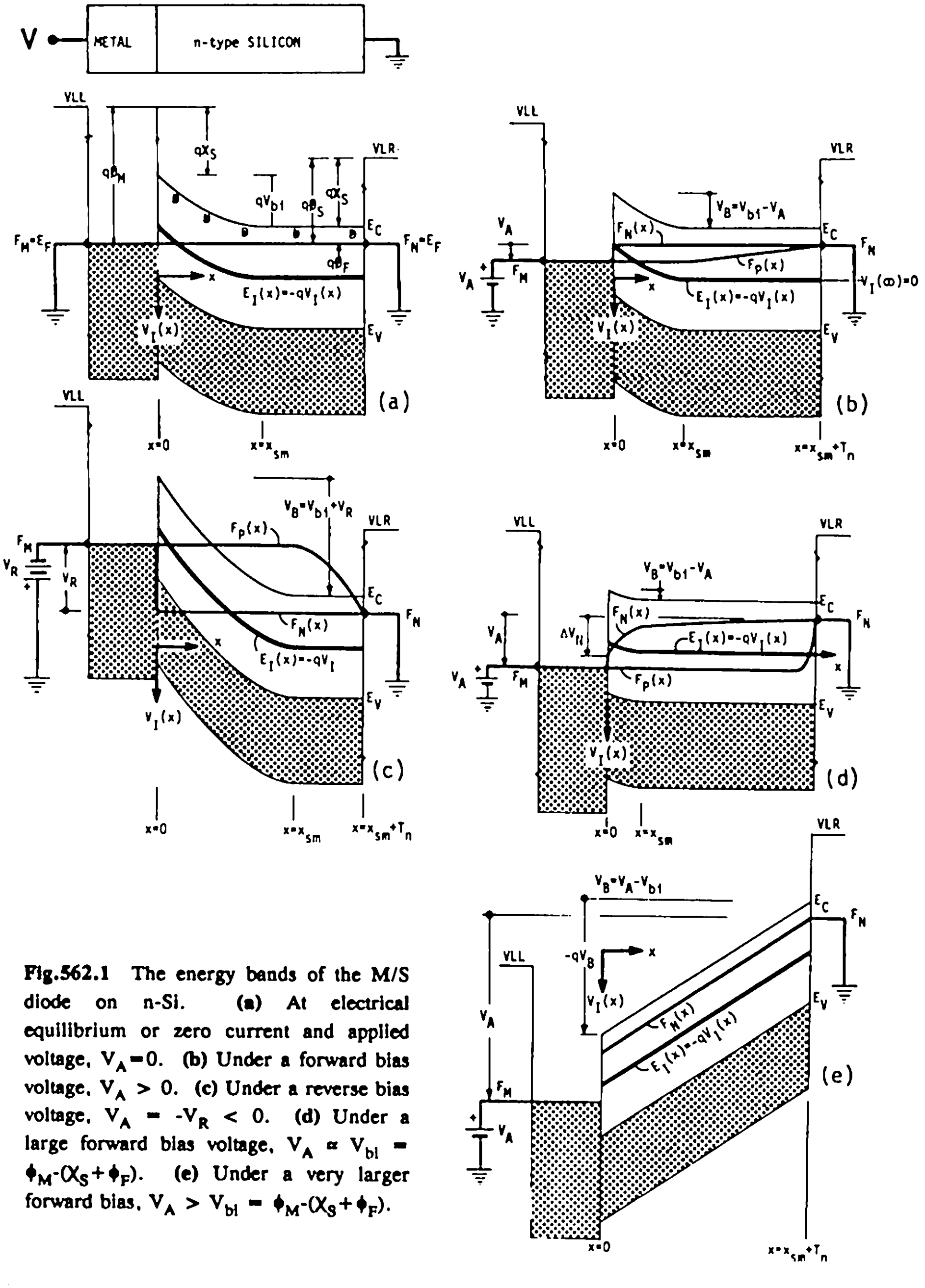

V
METAL
n-type SILICON
VLL
qØM
qXS
VLR
qVbi
qØS
qXS
EC
FM=EF
FN=EF
qØF
x
EI(x)=-qVI(x)
VI(x)
EV
(a)
x=0
x=xsm
VB=Vbi-VA
VA
FM
FN(x)
FP(x)
FN
-VI(∞)=0
(b)
x=xsm+Tn
FP(x)
VB=Vbi+VR
VR
FN(x)
EI(x)=-qVI
(c)
ΔVN
(d)
VB=VA-Vbi
-qVB
(e)

Fig.562.1 The energy bands of the M/S diode on n-Si. **(a)** At electrical equilibrium or zero current and applied voltage, $V_A = 0$. **(b)** Under a forward bias voltage, $V_A > 0$. **(c)** Under a reverse bias voltage, $V_A = -V_R < 0$. **(d)** Under a large forward bias voltage, $V_A \approx V_{bi} = \phi_M - (\chi_S + \phi_F)$. **(e)** Under a very larger forward bias, $V_A > V_{bi} = \phi_M - (\chi_S + \phi_F)$.

Recall that the electron current given by (562.7) is due to the electrons flowing from the semiconductor into the metal. Diode and transistor engineers have called this the **hot** electron current and the m/s diode the hot electron diode on the premise that the semiconductor electrons passing over the semiconductor barrier into the metal will have a kinetic energy of $q(\phi_M-\chi_S)$ or greater, and the metal electrons passing over the metal barrier into the semiconductor may have a kinetic energy of qV_B, giving an effective electron temperature of $T_e=q(\phi_M-\chi_S)/k$ and qV_B/k, or 11,605K per electron-volt. Actually, phonon scattering is so rampant in both the semiconductor and the metal that the energetic electrons will quickly lose their kinetic energy with the consequence that their energy distribution is essentially that at thermal equilibrium, the Fermi-Dirac or Boltzmann distribution. This is known as the **thermalization** of the hot electrons. Thus, the term, hot electron diode, is misleading and conceptually incorrect for the mass-produced m/s Schottky barrier diodes. In fact, in the earlier 1970s, it was even suggested that a metal base or hot electron transistor can be built to give exceedingly high frequency performance by using a forward biased S/M junction as the emitter to inject hot electrons (majority carriers) into a very thin metal-base, and using another and reverse-biased S/M junction on the other surface of the thin metal base to collect the injected hot electrons. The multi-year project undertaken at Bell Laboratories was terminated since no useful current gain could be obtained due to thermalization and a second factor, the quantum mechanical reflection of the electron wave passing over the s/m barrier of the collector junction. Thus, the current or thermionic current passing over the s/m barrier is the nearly thermal equilibrium current and is not limited by hot electron transport in the metal or the semiconductor. The exception is when the reverse bias voltage applied to the s/m diode reaches the current runaway or breakdown voltage due to interband impact generation of electron-hole pairs by hot electrons and holes in the semiconductor space-charge layer. Then, hot electrons and holes do exist in the semiconductor space-charge layer of the s/m diode. This is the hot electron effect described for p/n junction space-charge layer in section 536 and not the non-existent hot electron injection effect assumed by transistor engineers to coin the term, hot electron diode.

The net current crossing the metal/semiconductor interface at $x=0$ is then

$$J_N = \overrightarrow{J}_{Nx} + \overleftarrow{J}_{Nx} = J_{TH}\exp[-q(\phi_M-\chi_S)/kT]\exp(qV/kT) + \overrightarrow{J}_{Nx} \qquad (562.9)$$

To evaluate the current due to electrons flowing from the metal into the semiconductor, $J_{Nx}{}^{\leftarrow}$, we need not go through the evaluation of an integral similar to (562.7). A much simpler derivation can be made by applying the zero current condition at zero voltage to (562.9), $J_N(V=0) = 0$. This gives

$$\overrightarrow{J}_{Nx} = - J_{TH}\exp[-q(\phi_M-\chi_S)/kT] = J_{HB}. \quad \text{(Richardson Equation)} \qquad (562.9A)$$

which is independent of the applied voltage, V, because the metal barrier height, $\phi_M-\chi_S$, is independent of the applied voltage, unlike the semiconductor barrier height V_B, given by (562.1A)-(562.1C), which is proportional to the applied voltage. Equation (562.9A) is known as the **Richardson-Dushman** or **Richardson equation** of electrons emitted from a hot cathode in vacuum tube electronics.

Using (562.9A) in (562.9), the net current flowing through a m/s barrier is

$$J_N = \overleftarrow{J}_{Nx} + \overrightarrow{J}_{Nx} = J_{TH}\cdot\{\exp[-q(\phi_M-\chi_S)/kT]\}\cdot[\exp(qV/kT) - 1]$$
$$= J_{HB}\cdot[\exp(qV/kT) - 1]. \qquad (562.10)$$

This was derived by Hans Bethe in 1942 in a U.S. radar research effort for World War II at the MIT Radiation Laboratory (predecessor of the Lincoln Laboratory) and MIT Research Laboratory of Electronics. It will be known in this book as the **Bethe diode equation. Its saturation current**, J_{HB}, is given by

$$J_{HB} = J_{TH}\cdot\exp[-q(\phi_M-\chi_S)/kT]$$
$$= 120(m^*/m)T^2\cdot\exp[-q(\phi_M-\chi_S)/kT] \quad \text{A/cm}^2 \qquad (562.11)$$
$$= 10.8(m^*/m)(T/300)^2\exp[-q(\phi_M-\chi_S)/kT] \quad \text{MA/cm}^2 \qquad (562.11A)$$

where the subscript HB stands for Hans Bethe. The numerical constant 120, from $qmk^2/2\pi^2\hbar^3$ as indicated by (562.8), is known as the **Richardson constant** in vacuum electronics. Ancient experimental values obtained in vacuum electronics (listed in Kittel, p.247, 3rd-1966 edition which was deleted in the 6th-1986 edition) are smaller due to surface contamination by impurities resulting in lower surface barrier or metal work functions. The difference should be represented by ϕ_M, rather than absorbed into 120 using an empirical J_{HB}^* as practiced by some transistor engineering books and authors (see Sze, p.255 2nd-1981 edition). The equivalent $\delta\phi_M$ listed by these authors is so small that its physical significance cannot be ascertained.

At large reverse bias voltages, $V << -kT/q$, we have $\exp(qV/kT) << 1$ so that $\exp(qV/kT)$ can be neglected compared with 1 in $[\exp(qV/kT)-1]$ of the Bethe diode equation (562.10). Then, $J_N \approx -J_{HB}$ which shows that the Bethe diode current saturates to a voltage-independent constant value at large reverse bias voltages.

The Bethe diode equation, (562.10), has exactly the same form as the Shockley diode equation, (532.14). However, their coefficients or saturation currents, J_{HB} and J_1, are very different since their physical origin is entirely different: the Shockley diode is limited by diffusion and recombination of minority carriers injected into or extracted out of the quasi-neutral emitter and base layers of the p/n or other junctions; while the Bethe diode is limited by thermionic emission of the majority carriers over the potential barrier. In fact, the Bethe diode saturation current (or thermionic saturation current) is substantially larger than the Shockley diffusion-recombination saturation current,

$$J_{HB} >> J_1 = J_{p0} + J_{n0}. \qquad (562.12)$$

One can readily prove this numerically. From the numerical data of the p+/n junction diode calculated in section 534 we had $J_{p0}=4.8\text{x}10^{-10}$ A/cm^2 in (534.3a) and $J_{n0}=8.0\text{x}10^{-14}$ A/cm^2 in (534.3b). Thus,

$$J_1 = J_{p0} + J_{n0} \propto J_{p0} = 4.8\text{x}10^{-10} \text{ A/cm}^2.$$

For a Bethe diode made by a metal such as Al on n-type Si, the experimental equilibrium barrier height from the metal to the semiconductor is about 0.68 eV. The Bethe diode saturation current is then

$$\begin{aligned} J_{HB} &= 120(m^*/m)T^2\cdot\exp[-(\phi_M-X)/kT] \\ &= 120\text{x}1\text{x}300^2\cdot\exp[-0.68/0.02585] = 4.064\text{x}10^{-5} \text{ A/cm}^2. \end{aligned} \qquad (562.13)$$

This is five orders of magnitude larger than the Shockley diode saturation current. Thus, there is essentially no minority carrier injection and hence no minority carrier charge storage in a Bethe diode. This enables Schottky-barrier Bethe diodes to turn off at very high speed.

One can also demonstrate numerically that the recombination-generation current in the space-charge layer of a p/n junction diode is several orders of magnitude smaller than the Bethe diode current. Experimental examples are given in the next section.

563 Experimental M/S Diodes

The d.c. forward and reverse current-voltage characteristics of four metal/n-Si Bethe (Schottky-barrier) diodes and three historical diffused Si p+/n junction Shockley diodes of 1957, 1962 and 1977 vantages are compared in Figs.563.1(a) and (b). The metals, tungsten (W), aluminum (Al), and tungsten-disilicide (WSi$_2$) are three of the commonly used contact metals in silicon integrated circuits. Others used in current integrated circuit technologies include Ti, Ta and their silicides. They are selected for process compatibility so that the very thin (100-1000A) and very small (submicron) oxides and diffused junctions are not destroyed by these metals and their silicides at high oxidation and diffusion temperatures and during alloying to give good ohmic contacts and strong mechanical adherence.

Figure 563.1(a) shows that the W/n-Si and Al/n-Si diodes have higher forward current densities than the WSi$_2$/n-Si and PtSi/n-Si diodes. This is due to larger work function difference and the presence of interface traps explained in the following paragraphs.

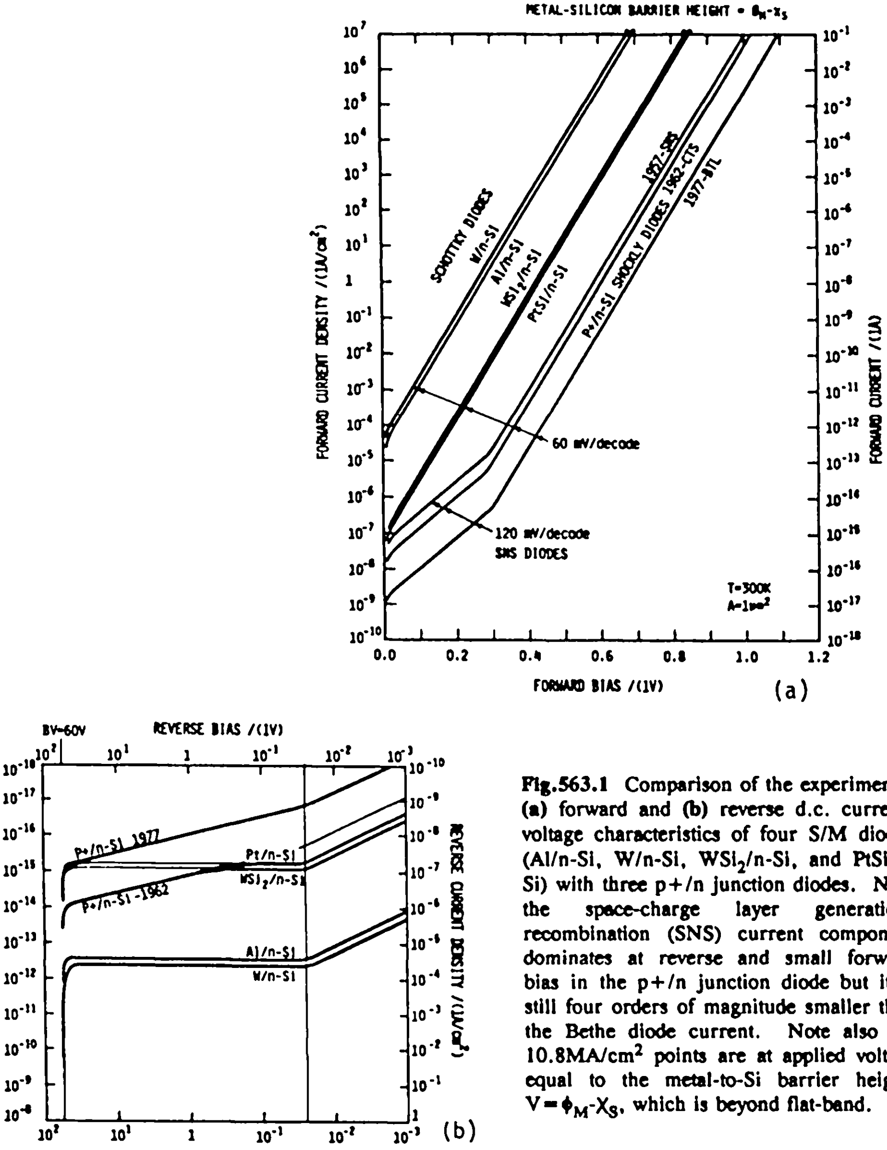


Fig.563.1 Comparison of the experimental (a) forward and (b) reverse d.c. current-voltage characteristics of four S/M diodes (Al/n-Si, W/n-Si, WSi$_2$/n-Si, and PtSi/n-Si) with three p+/n junction diodes. Note the space-charge layer generation-recombination (SNS) current component dominates at reverse and small forward bias in the p+/n junction diode but it is still four orders of magnitude smaller than the Bethe diode current. Note also the 10.8MA/cm^2 points are at applied voltage equal to the metal-to-Si barrier height, $V=\phi_M-\chi_S$, which is beyond flat-band.

The higher current densities is due to the smaller potential barrier heights of W/n-Si and Al/n-Si than WSi$_2$/n-Si and PtSi/n-Si. The Bethe diode leakage current density, J_{HB} given by (562.11), is smaller if the metal work function, ϕ_M, is larger since the metal barrier height is given by $\phi_M-\chi_S$. For silicon, the electron affinity is $\chi_S(\mathrm{Si})=4.029\mathrm{eV}$. The experimental work functions of the metals in vacuum are: (Al)4.2eV, (Pt)5.3eV, (Au)5.22eV, (W)4.565eV. If these vacuum numbers are used to calculate the metal barrier height, ϕ_M-χ_S, and the corresponding Bethe diode saturation current, J_{HB} given by (562.11), the computed results are substantially different from the experimentally observed values. For Al, the computed barrier height is ϕ_M-χ_S=4.2-4.029=0.171eV which is substantially smaller than the experimental value of 0.68eV. For Au, the computed barrier height is ϕ_M-χ_S=5.22-4.029=1.19eV which is significantly larger than the experimental value of 0.795±0.01eV. Discrepancies are also observed for other metals, such as Ag(4.31), Cu(4.52), Pd(5.0), whose computed and observed barrier height, ϕ_M-χ_S, are Ag(0.28, 0.655±0.01), Cu(0.49, 0.550±0.03) and Pd(0.97, 0.710±0.005).

The discrepancy is due to the idealized or perfect metal/semiconductor interface assumed in our derivation of the Bethe diode current-voltage equation, (562.10). In practice, there is always a thin and nonuniform layer of residual silicon dioxide between the metal and the silicon. The thin oxide layer creates a high density of interface traps whose energy levels are distributed both in energy throughout the Si energy gap and in space into the oxide from the SiO_2/Si interface. Interface trap was called 'surface state' by Tamm(1939), Shockley(1939) and Bardeen(1946) and in the older literature and textbooks. The high density of interface traps at the metal/thin-residual-oxide/Si interface is caused by a deficiency of oxygen atoms to tie up all the silicon bonds at the oxide/silicon interface when the Si is oxidized either in a high-temperature oxidation furnace or at room temperature and in room air. There are about $5\text{x}10^{14}$ cm^{-2} of Si-O bonds at the SiO_2/Si interface. Interface trap density of 10^{13} cm^{-2} can be produced in thermally oxidized Si and $4\text{x}10^{14}$ cm^{-2} is observed on vacuum cleaved Si surface where all the Si bonds are broken and there is no residual oxygen-containing gas in the vacuum bell jar. In VLSI MOS chip production, the SiO_2/Si interface trap density has been routinely reduced to $10^9 cm^{-2}$ or less by clean oxidation and proper cooling and annealing described in chapter 4 on MOS capacitors. This corresponds to only about one broken bond among 10^6 unbroken or good Si-O bonds. However, the processing conditions of the VLSI metal/Si contacts and Schottky barriers do not give significant reduction of interface traps at a metal/Si contact.

These broken or dangling silicon bonds are electron and hole traps. When an interface trap captures an electron it becomes more negatively charged and when it captures a hole, it becomes more positively charged. Because of the very high density of interface traps in the metal/Si interfacial layer, the Si surface potential barrier height at the interface will adjust itself to a value in order to keep the net total interface trapped charge nearly zero. This is effected by bending the Si energy band at the interface until half of all the interface traps, located in the lower part of the Si energy gap, are occupied by electrons and become electrical neutral (since these lower-energy interface traps are donor-like), and the other half, located in the upper part of the Si energy gap, are unoccupied by electrons or occupied by holes and also become electrical neutral (since these higher-energy interface traps are acceptor-like). Because of the very high density of the interface traps, the foregoing **charge-neutral condition** on the occupation of the interface traps by electrons and holes overwhelmingly determines the semiconductor space-charge density. Thus, it determines the position of the Fermi level at the metal/Si interface and the amount of semiconductor energy band bending or the semiconductor built-in barrier height. Band bending from the metal/semiconductor work function difference is negligible compared with that due to the interface traps just described. This Fermi level position at the oxide/Si interface or a barely oxide-covered Si surface is known as the neutral-trap Fermi level or neutral Fermi level, E_N.

The pinning or locking of the Fermi level to the neutral Fermi level position at the metal/semiconductor interface not only causes the experimental Schottky barrier height to differ from that calculated using the vacuum work function value of the metal but also makes the Si surface band bending or barrier height nearly independent of the type of metal or conductor used for the metal/Si Schottky-barrier Bethe diodes.

The foregoing description of the physics of the metal/semiconductor potential barrier height was advanced by John Bardeen in his famous classic article of 1947 on surface states (or interface traps) on metal/semiconductor diodes published in the Physical Review [563.1].

[563.1] John Bardeen, "Surface states and rectification at a metal semiconductor contact," Physical Review, 71(10) 717-727, May 15, 1947.

It was this understanding of the basic physics of the surface and interface that led to the invention of the bipolar junction transistor and the junction-gate field-effect transistor by Shockley who published the transistor design theories in 1949 and 1951. It was also this fundamental understanding that led eventually to the successful development and volume production of stable and low interface-trap-density silicon MOS field-effect transistors at the Fairchild Semiconductor Corporation during 1959 to 1962, and the volume production of inexpensive and reliable silicon MOS integrated circuits such as the DRAM (dynamic random access memory) and MPU (multi-processor unit used in the IBM PC's and PS-2 personal computers) by the Intel Corporation beginning in the late 1970's.

Returning to Figs.563.1(a) and (b), all the diodes there have $N_{DD} \approx 10^{16}\text{cm}^{-3}$ which gives a reverse breakdown voltage of 60V as indicated in figure (b). The carrier lifetimes of the three p+/n diodes range from 0.1 to 100μs, the 1977 Bell Telephone Laboratory technology had the longest. It is evident that the W/n-Si and Al/n-Si Schottky-barrier Bethe diodes have the highest reverse leakage current. However, the leakage currents of the WSi_2/n-Si and PtSi/n-Si Schottky-barrier Bethe diodes are similar to that of the 1962 p+/n Shockley diode whose reverse current was dominated by the SNS current due to the thermal generation of electrons and holes from the generation centers in the space-charge layer. Thus, if the surface space-charge layer of the WSi_2/n-Si and the PtSi/n-Si Schottky diodes contains a generation center at the same concentration as the 1962 Si p+/n diode, then the reverse current of these Schottky barrier diodes would contain two components: the Bethe thermionic current over the Schottky barrier, and the SNS thermal generation current from the traps in the Si surface space-charge layer. The total current of such a Schottky-barrier diode would be given by the composite dark curve in the upper middle part of Fig.563.1(b) whose left side at higher voltages is labeled P+/n-Si (1962) and whose right side at lower voltages is labeled WSi_2/n-Si.

The main conclusion of this comparison is that Schottky barrier diodes pass considerably higher current or current density at a given forward and reverse bias voltage than a comparable p+/n junction diode.

564 Effect of Semiconductor Voltage Drop - Mott Theory

At high forward current densities, the majority carriers (electrons) flowing in the semiconductor (n-Si) space-charge and quasi-neutral bulk layers will cause voltage drops which reduce the junction voltage and current. These voltage drops are measured by the total change of the electron (majority carrier) quasi-Fermi potential across the space-charge and quasi-neutral bulk layers. The voltage drops would reduce the separation of the quasi-Fermi potentials at the metal/n-Si interface, thereby increasing the silicon barrier height, and reducing the forward current. This is the **ohmic voltage drop** effect but it has been known to device theorists, engineers and book authors as the diffusion theory which is misleading and a misnomer since the theory includes both the drift and the diffusion currents and not just the diffusion current alone. The use of the single term, diffusion, has caused considerable confusion on its true physical origin. The result of this diffusion-drift theory will be known as the **Mott diode equation** of a Schottky barrier diode, the **Mott Schottky-barrier diode equation**, or just the **Mott diode equation** after Mott who derived it first in 1939. Mott's theory was the first to predict the correct rectification direction but not the magnitude. Bethe predicted the correct magnitude as well as the rectification direction. Other previous theories had predicted the wrong direction of easy current flow as well as magnitude, some of which invoked tunneling which has become technologically important for low-resistance ohmic contacts described in section 583.

The following paragraphs describe the derivation of the Mott diode equation by taking into account the voltage drop. The total voltage drop through the space-charge and quasi-neutral layers is given by [See Fig.562.1(d).]

$$\Delta V_N = V_N(x=0) - V_N(x=x_{ms}+T_n) = V_A - V_J, \tag{564.1}$$

V_A is the applied terminal voltage. V_J is the junction voltage or the quasi Fermi potential difference at the semiconductor side of the interface, $V_P(0)$-$V_N(0)$. This was denoted by V without the subscript J in the Bethe diode equation, (562.10). Subscript J is now used to distinguish the applied terminal voltage, V_A, and the voltage at the junction, V_J, which is smaller due to the voltage drops. x_{ms} is the thickness of the space-charge layer and T_n is the thickness of the quasi-neutral n-type layer. The voltage drop and the junction voltage V_J, can be readily computed by integrating the electron current density through the space-charge and quasi-neutral layers since J_N is spatially constant owing to the assumption of negligible recombination-generation currents depicted by the d.c. form ($\partial/\partial t=0$) of the two continuity equations (350.1) and (350.2) when $g=r=0$. From (331.1), the d.c. electron current is given by

$$J_N = -q\mu_n N dV_N/dx = J = \text{spatially constant.} \tag{564.2}$$

Denote the total change of the majority carrier (electron) quasi-Fermi potential by ΔV_N as indicated by the high-current energy band diagram shown in Fig.562.1(d) and call this the voltage drop, then

$$\Delta V_N = V_N(0) - V_N(x_{ms}+T_n) = \int_0^{x_{ms}+T_n} Jdx/[q\mu_n N(x)] \tag{564.3}$$

$$\geq [(x_{ms}+T_n)/(q\mu_n N_{DD})]J \tag{564.3A}$$

$$= [R_{ms} + R_{nb}]J. \tag{564.3B}$$

Equation (564.3B) shows explicitly the voltage drops across the two series resistances, R_{ms} and R_{nb}, where R_{ms} is resistance (or majority carrier resistance) of the space-charge layer defined by

$$R_{ms} \geq x_{ms}/(q\mu_n N_{DD}) \tag{564.4A}$$

and R_{nb} is the majority carrier resistance of the quasi-neutral n-layer defined by

$$R_{nb} = T_n/(q\mu_n N_{DD}). \tag{564.4B}$$

To be complete, the voltage drop across the resistance of the ohmic contact to the back surface of the semiconductor, R_{nc}, should be added to the resistances in (564.3B).

The approximation in (564.3A) and (564.4A) comes from the assumption of $N(x)=N_{DD}$ in the space-charge layer which underestimates the resistance of the space-charge layer since $N(x)<N_{DD}$ in the space-charge layer. This reduction of the electron concentration in the space-charge layer is due to the transfer of electrons from the semiconductor to the metal owing to the larger metal work function than semiconductor.

The complete S/M diode equation must include the voltage drop across these resistances. They are included by replacing the junction voltage V in the Bethe diode equation, (562.10), with V_J or V_A-ΔV_N. Rewrite (562.10) using $V=V_J=V_A-\Delta V_N$, then the generalized Bethe diode equation is

$$J = J_{HB}[\exp(qV_J/kT) - 1] \qquad (564.5)$$

$$= J_{HB}\{\exp[q(V_A-\Delta V_N)/kT] - 1\}. \quad \text{(Bethe Diode)} \qquad (564.6)$$

Using the under-estimated resistance approximation, (564.3A), to compute the voltage drop $\Delta V_N \gtrsim (R_{ms}+R_{nb})J$, then the Bethe diode current from (564.6) is

$$J \simeq J_{HB}\cdot\left[\exp\{q[V_A - J(x_{ms}+T_n)/(q\mu_n N_{DD})]/kT\} - 1\right]. \qquad (564.7A)$$

This is transcendental in J for a given applied voltage V_A. It can be solved implicitly for the diode current at a given applied voltage only by numerical iteration. However, it may be solved explicitly for the terminal voltage at a given applied current without numerical iteration. The explicit equation is obtained by inverting (564.7A)

$$V_A = (kT/q)\log_e[(J/J_{HB}) + 1] + (R_{ms} + R_{nb} + R_{nc})J$$

$$= (kT/q)\log_e[(J/J_{HB}) + 1] + \{[(T_n+x_{ms})/(q\mu_n N_{DD})] + R_{nc}\}J \qquad (564.7B)$$

which enables us to compute the diode terminal voltage, V_A, for a given applied diode current density J without numerical iteration.

The exact solution of the Mott diode current can be obtained using

$$N(x) = n_i\exp\{+[F_N(x)-E_I(x)]/kT\}$$

$$= N_N\exp\{(q/kT)[V_I(x)-V_N(x)]\}$$

$$\simeq N_{DD}\exp\{(q/kT)[V_I(x)-V_N(x)]\} \qquad (564.8)$$

to integrate (564.2) over the space-charge and quasi-neutral layers. Different reference potentials are used for $V_I(x)$ and $V_N(x)$ as illustrated in Fig.562.1(d): $V_I(x=x_{ms}+T_n)=0$ and $V_N(x_{ms}+T_n)=0$. This integration gives

$$J = J_N = \frac{qD_nN_{DD}[1 - \exp(-q\Delta V_N/kT)]}{\int_0^{x_{ms}+T_n} \exp[-qV_I(x)/kT]dx}. \quad \text{(Mott Diode)} \qquad (564.9)$$

We shall call this equation the **Mott diode equation** in this book. It can be combined with (564.6) to eliminate the voltage drop, ΔV_N. The final and complete equation of the S/M diode is given by

$$J = \frac{J_{HB}[\exp(qV_A/kT) - 1]}{1 + \dfrac{J_{HB}\exp(qV_A/kT)}{qD_nN_{DD}} \displaystyle\int_0^{x_{ms}+T_n} \exp[-qV_I(x)/kT]dx} . \tag{564.10}$$

This exact solution can be simplified to give an approximate analytical solution if the depletion approximation is used to give the quadratic variation of the electric potential through the space-charge layer and if the constant electric field approximation is used to give the linear variation of the electric potential in the n-type quasi-neutral layer. This derivation is left as an exercise. Instead, we shall demonstrate that this exact result reduces to the correct asymptotic solutions at low and high currents which can be written down immediately.

At low current densities when J or $J_{HB}\exp(qV_A/kT) << qD_nN_{DD}/(x_{ms})$, the second term in the denominator of (564.10) is much smaller than 1. Thus, (564.10) reduces to the Bethe diode equation given by the numerator which is

$$J = J_{HB}[\exp(qV_A/kT) - 1]. \quad \text{(Low-Current Bethe diode)} \tag{564.11}$$

At high current densities when $V_A \geq V_{bi}(=V_D)$ and $J_{HB}\exp(qV_A/kT) >> qD_nN_{DD}/(x_{ms}+T_n)$, the second term in the denominator of (564.10) is much greater than 1. Thus, (564.10) reduces to the resistance equation

$$J \propto \frac{qD_nN_{DD}}{\displaystyle\int_0^{x_{ms}+T_n} \exp[-qV_I(x)/kT]dx} = \frac{qD_nN_{DD}}{\displaystyle\int_0^{x_{ms}+T_n} \exp[-qV_I(x)/kT](dx/dV_I)dV_I} \tag{564.12}$$

$$\leq \frac{q_nN_{DD}[-dV_I(0)/dx]}{\displaystyle\int_0^{x_{ms}+T_n} \exp[-qV_I(x)/kT]dV_I} = \frac{qD_nN_{DD}[-dV_I(0)/dV]}{\{1 - \exp[-q\Delta V_I(0)/kT]\}(kT/q)}$$

$$\leq q\mu_nN_{DD}(V_A-V_{BI})/(x_{ms}+T_n) = (V_A-V_{BI})/(R_{ms}+R_{nb}). \tag{564.12A}$$

This is the resistance equation of a piece of semiconductor of length $x_{ms}+T_n$ with a built-in battery of V_{bi}. This built-in battery comes from the metal/semiconductor work function difference which gives the built-in or diffusion potential of the metal/semiconductor barrier at zero current and voltage, $V_{bi}=\phi_M-\phi_S=\phi_M-(\chi_S+\phi_F)$. The energy band diagram at this high forward voltage is given in

Fig.562.1(e) which shows the constant electric field throughout the entire semiconductor due to the resistive voltage drop through the semiconductor.

These two asymptotic analytical solutions prove that the historical notion of the diffusion theory proposed by Wagner in 1931, Mott in 1939, and Schottky and Spenke in 1939, for the metal/semiconductor diode and held by previous and recent authors of journal articles and textbooks is erroneous and incomplete. It is not just the diffusion mechanism as the term 'diffusion theory' would suggest. The derivation just described explicitly included the drift current which actually dominates (smaller than and in series with) the thermionic or Bethe diode current and limits the diode's terminal current at high forward current densities. The drift current, when dominant, simply gives the voltage drop across the integrated majority carrier resistance through the n-type semiconductor. In his 1942 thermionic theory, Bethe recognized the fallacy of the diffusion-dominance argument used in the 1939 Mott-Schottky theory. Bethe made the assumption that the scattering mean-free-path must be large compared with the thickness of the space-charge layer in order to invalidate the diffusion theory and in order to have the thermionic current dominant over the diffusion current. Such a one-dimensional comparison of the mean-free-path with the thickness of the space-charge-layer overlooks the scattering events in the transverse directions. However, none of them recognized the resistance voltage-drop aspect presented here that is imbedded in the Mott-Schottky-Davydov diffusion-drift theory due to their preoccupation by diffusion.

565 Integrated Circuit Schottky Barrier Diode Layouts

Figure 565.1(a) gives the cross-sectional view of a most compact practical design of an S/M (Schottky barrier) diode that is in parallel with a p+/n junction diode. The metal electrode (dark layer) forms the ohmic contact with the p+ diffused region of the p+/n diode on the left side of the cross sectional view. It also serves as the metal of the metal/n-Si Bethe diode on the right side of the cross sectional view. The circuit symbol of the M/S Schottky-barrier Bethe diode is shown in Fig.565.1(b) which is our preference. A normal boxed shaped S, for Surface-barrier, is indicated in contrast to some usages which have employed inverse S.

Figure 565.2 shows one application, perhaps the most important and certainly the largest volume application of the Bethe-Schottky barrier diode. It is connected across the collector base junction of a bipolar junction transistor to prevent the transistor from going into saturation which would forward bias the collector junction. This S/M diode minimizes the injection of minority carriers by the collector junction which would occur in heavy saturation since the injected charge would have to be stored in the collector and base layers. The stored charge would delay and extend the switch-off transient and increase the switching time. Fig.565.2(a) shows an integrated-circuit layout of a Bethe-Schottky-diode-clamped n+/p/n/n+ bipolar junction transistor (BJT). Fig.565.2(b) shows the separate symbols for the Bethe (Schottky) diode and the n/p/n BJT. Fig.565.2(c) is the more compact circuit symbol used for a Bethe(Schottky)-diode-clamped BJT which must be qualified by noting that the Bethe (Schottky) diode clamp is connected across the collector-base junction only and not the emitter-base junction. The latter would have reduced the BJT current gain to zero.

Note again that the —▶ʃ— symbol contains the conventional diode symbol and an 'S' which stands for the Schottky or Surface barrier. A more appropriate symbol is due Bethe who gave the correct theory in 1942, such as —▶B— .

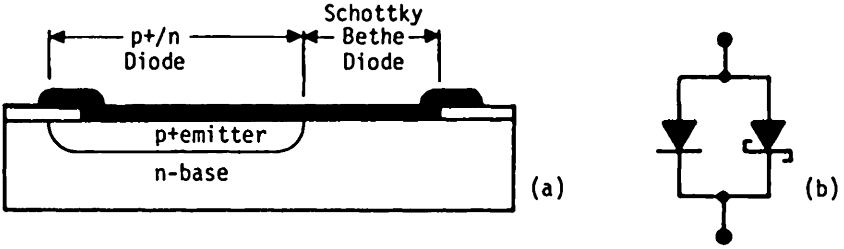

Schottky
p+/n Diode
Bethe Diode
p+emitter
n-base
(a)
(b)

Fig.565.1(a) Integrated circuit Bethe (Schottky-barrier) diode in parallel with a p+/n diffused junction diode. (b) Circuit symbols for the Bethe (Schottky) diode and the p+/n diode.

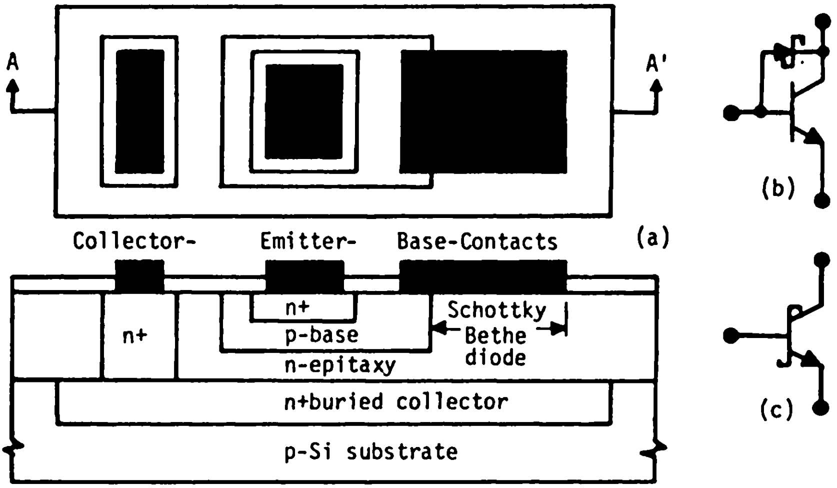

A
A'
Collector-
Emitter-
Base-Contacts
(a)
n+
n+
p-base
Schottky
Bethe
diode
n-epitaxy
n+buried collector
p-Si substrate
(b)
(c)

Fig.565.2(a) The top and cross sectional views of an integrated Bethe (Schottky-barrier) diode with a n/p/n BJT (Bipolar Junction Transistor) used in TTL (Transistor-Transistor Logic) circuits. ((b) Circuit symbols of separate BJT and SBD. (c) Circuit symbol of the (Bethe) Schottky-barrier diode clamped bipolar junction transistor.

570 TUNNEL DIODES

Tunneling was first proposed by Zener in 1934 as a possible mechanism for dielectric breakdown. It cannot give electronic breakdown since there is only one carrier species and no regenerative feedback from a second species. Tunnel current in reverse biased p/n junction diode was first identified by K.G. McVay in 1950 at Bell Labs (see section 287), and in forward biased p/n junction by Leo Esaki in 1957 at Sony. The latter attracted tremendous interest because its negative resistance has no theoretical frequency limit. Fig.570.1(a) shows Sah's 1961 experimental forward I-V curves of Si p+/n+ tunneling diodes with increasing gold density for studying the excess current. Tunnel transition paths via Au-trap are shaded in figure (b). The interband electron tunnel current (dark arrow) can be calculated like the Bethe current of section 562 with two modifications: an attenuation-transmission factor in the forbidden path given by (154.1) and the final density of unoccupied state. From (562.2) the tunnel current is

$$\overleftarrow{J}_{Nx} = q\int v_x dn(1-f_{P+})e^{-2\Theta} = q\int(\partial E/\hbar\partial k_x)f_{N+}(1-f_{P+})e^{-2\Theta}[2dk_x dk_y dk_z/(2\pi)^3]$$

The exponent of the attenuation factor is the barrier integral $\Theta = \int k_x dx$. k_x is given by $Fx - E = (\hbar^2/2)[(k_x^2/m_x)+(k_y^2/m_y)+(k_z^2/m_z)]$. Analytical solution can be obtained only at very low temperatures when the Fermi function can be approximated by the unit-step function. Nevertheless, the formulae predicts the observed tunnel current accurately. The negative resistance comes from the decreasing overlap of f_{N+} (electrons or occupied states) on the n^+ side with $1-f_{P+}$ (holes or unoccupied states) on the p^+ side. If a triangular barrier or a constant electric field is assumed in the p^+/n^+ junction, then the Fowler-Nordheim formulae given in section 385 and Fig.385.1 can be used to estimate the tunneling current at low voltages with electric field scaled by $(E_G/\Phi_B)^{3/2} = (1.2\text{eV}/3.1\text{eV})^{3/2} = 0.24$.

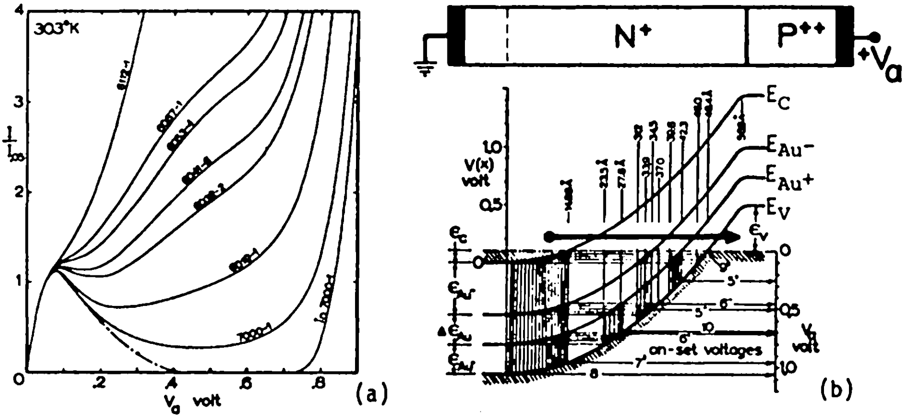


Fig.570.1 Si p^+/n^+ tunnel diode. (a) Experimental I-V with Au-doping. (b) Energy band diagram.

580 LIMITING MECHANISMS OF DC TERMINAL CURRENT OF DIODES

The preceding analyses show that there are five electric current components or mechanisms in the **terminal** current of p/n and m/s diode structures. These are the diffusion and drift currents of majority and minority carriers, and the recombination-generation current of the two carriers in the junction space-charge layer. Each of the five current components can limit the terminal current in a particular diode structure depending on the applied d.c. voltage and d.c. current density. For example, at medium-to-high current densities when there is still a potential barrier of more than several kT/q's separating the two sides, the terminal current in a Shockley p/n junction diode is limited by minority carrier diffusion-recombination-generation in the quasi-neutral layers; while in a m/s Bethe diode, it is limited by thermionic emission of majority carriers over the potential barrier. An analysis of the limiting mechanisms of the terminal current in m/s and p/n diodes at high current levels are given in the following two sections, 581 and 582. At very high forward current densities, the contact resistance may become limiting which is analyzed in section 583.

581 Current Limits in Metal/Semiconductor Diodes

The **terminal** current of metal/semiconductor diodes is limited by the **majority carrier thermionic current** (Bethe diode) flowing over the potential energy barrier or Schottky barrier of the m/s junction at low current densities. At intermediate and higher current densities when the forward d.c. bias voltage approaches the built-in semiconductor potential barrier height, the terminal current is mainly limited by the voltage drop across the **majority carrier diffusion and drift resistances** in the space-charge and quasi-neutral-bulk layers of the semiconductor (Mott diode) if the contact resistances are small. At still higher forward current densities when the applied voltage substantially exceed the built-in potential barrier height, the terminal current becomes completely limited by the **majority carrier drift current** and **resistance** of the quasi-neutral-bulk layer and the contact resistances. The minority carrier diffusion and drift currents are negligible at all current levels.

The maximum thermionic current density (not the terminal current density which can be higher than the maximum thermionic current density) that can pass through the Schottky barrier occurs when the semiconductor barrier height given by (562.1B) is reduced to zero by the forward applied voltage: $0 = V_B = V_{bi} - V_A = \phi_M - \phi_S - V_A = \phi_M - (\chi_S + \phi_F) - V_A$. This is the **flat band** condition. It occurs when the forward applied voltage equals the built-in semiconductor barrier height, $V_A = V_{bi} = \phi_M - \phi_S = \phi_M - (\chi_S + \phi_F)$. Above this current density, the Schottky potential barrier no longer exists and the bulk resistance of the semiconductor limits the terminal current density if the contact resistance is negligible. Substituting the above applied voltage into (562.10) and using (562.11) for the numerical values in J_H, this maximum thermionic current density is given by

$$\begin{aligned}
J &= J_{HB}[\exp(qV_A/kT) - 1] \simeq J_{HB}\exp(qV_A/kT)\\
&= J_{HB}\exp[q(\phi_M-\chi_S-\phi_F)]\\
&= J_{TH}\exp[-q(\phi_M-\chi_S)/kT]\exp[q(\phi_M-\chi_S-\phi_F)/kT] = J_{TH}\exp(-q\phi_F/kT)\\
&= 120(m^*/m)T^2\exp(-q\phi_F/kT)\ \mathrm{A/cm^2}\\
&= 10.8(m^*/m)(T/300)^2\exp(-q\phi_F/kT)\ \mathrm{MA/cm^2}\\
&= q(\theta_{ave}/4)N \qquad (581.1)
\end{aligned}$$

Here N is the electron density in the n-type semiconductor and θ_{ave} is the average speed of the Boltzmann distributed electrons defined by

$$\begin{aligned}
\theta_{ave} &= \sqrt{(8/\pi)(kT/m^*)}\\
&= 1.076\times10^7(T/300)^{1/2}(m/m^*)^{1/2}\ \mathrm{cm/s} \qquad (581.2)
\end{aligned}$$

Plug in this number, then the limiting thermionic current density is

$$\begin{aligned}
J = q(\theta_{ave}/4)N &= 1.602\times10^{-19}\times(1.076\times10^7/4)\sqrt{(T/300)(m/m^*)}N\\
&= 4.309\times10^{-13}N\ (\mathrm{A/cm^2}) \qquad (581.3)
\end{aligned}$$

where N is the electron concentration in the n-type semiconductor in (#/cm^3) and where T=300 and $m^*=m$ are assumed. Note $J=10.8\mathrm{MA/cm^2}$=maximum thermionic current when $N=N_C=2.509\times10^{19}\mathrm{cm^{-3}}$ or $\phi_F=0$ and $E_F=E_C$, but the result is invalid because the Boltzmann approximation is poor when $\phi_F<2$.

In contrast to the average speed in the above result, the rms (root-mean-square) velocity (or speed) of the electrons defined by $m^*\theta_{rms}{}^2=3kT/2$ is

$$\begin{aligned}
\theta_{rms} &= \sqrt{3(kT/m^*)}\\
&= 1.168\times10^7(T/300)^{1/2}(m/m^*)^{1/2}\ \mathrm{cm/s} \qquad (581.4)
\end{aligned}$$

which is only 1.168/1.076=1.086 or 8.6% higher than the average velocity. Thus, in many instances, authors do not distinguish θ_{ave} and θ_{rms}. However, it is evident that one cannot use the average speed or the rms velocity to compute the thermionic current density since the effective velocity that gives the thermionic current density is only one-fourth of the average speed.

Does the thermionic current density limit exceed the scattering-limited drift velocity at high electric fields? If it does, then the scattering-limited drift current flowing in the n-type semiconductor would limit terminal current as the applied voltage is increased above the built-in barrier height. For Si, the scattering-limited velocities at T=300K are $\theta_n=1.0\times10^7$ cm/s for hot electrons and $\theta_p=0.75\times10^7$ cm/s for holes as indicated in Fig.314.1. Both of these scattering-limited velocities exceed one-fourth of the average velocity which is about 0.27×10^7 cm/s. Thus, the terminal current is limited by thermionic current in the entire current density range up to the maximum thermionic current density given by (581.3) at a given carrier concentration N which is approximately given by the dopant donor impurity concentration N_{DD} ($N\approx N_{DD}$) as demonstrated in chapter 2 when donor impurity deionization is unimportant ($N_{DD}<10^{18}cm^{-3}$). Above this thermionic current limit, the electron drift velocity will exceed one-fourth of the average thermal velocity and the thermionic current is no longer limiting the terminal current since the semiconductor potential barrier no longer exists to prevent the semiconductor electrons from moving into the metal. The series bulk resistance of the semiconductor, dominated by the majority carriers (electrons in the metal/n-Si diode), will limit the terminal current which is given by the limiting form of the Mott-diode current equation, (564.12A), of the Schottky barrier.

582 Current Limits in p/n Junction Diodes

The terminal current in the silicon p/n junction diode is limited by carrier recombination-generation in the space-charge layer under reverse bias or below the forward injection threshold voltage, V_{IT}, defined by (533.7). Above V_{IT}, the terminal current is limited by minority carrier diffusion-recombination in the two quasi-neutral layers at low forward current densities or low injection levels, and by ambipolar (combined majority and minority carrier) diffusion-recombination in the two quasi-neutral layers at high injection level. At higher current densities or injection levels when the carrier (electron and hole) densities greatly exceed the ionized dopant impurity densities, carrier-space-charge-limited ambipolar diffusion-recombination will limit the terminal current. At still higher current or injection levels, drift current and ohmic resistance of the two quasi-neutral layers and the two contacts will limit the terminal current.

Some of these limiting mechanisms may not occur in a given p/n junction diode. Their presence depends on the concentration of the dopant donors, dopant acceptors, and the recombination-generation centers, as well as the thickness of the layers. For example, the carrier-space-charge-limited ambipolar diffusion-recombination mechanism may not occur at the very high current densities if the majority carrier drift current or majority carrier resistance of one of the two quasi-neutral layers is already limiting the terminal current. Such a situation occurs in the thick quasi-neutral n-base (or thick wafer) of a p+/n diode.

The effect of voltage drop across the majority carrier resistance in the quasi-neutral and space-charge layers on the terminal current can be analyzed using the Mott-Schottky-Davydov approach. A modification is necessary since the Mott-Schottky-Davydov diode current is due to majority carrier drift and diffusion while in the p/n junction diode, it is due to minority carrier diffusion. This modification is described in the following paragraphs.

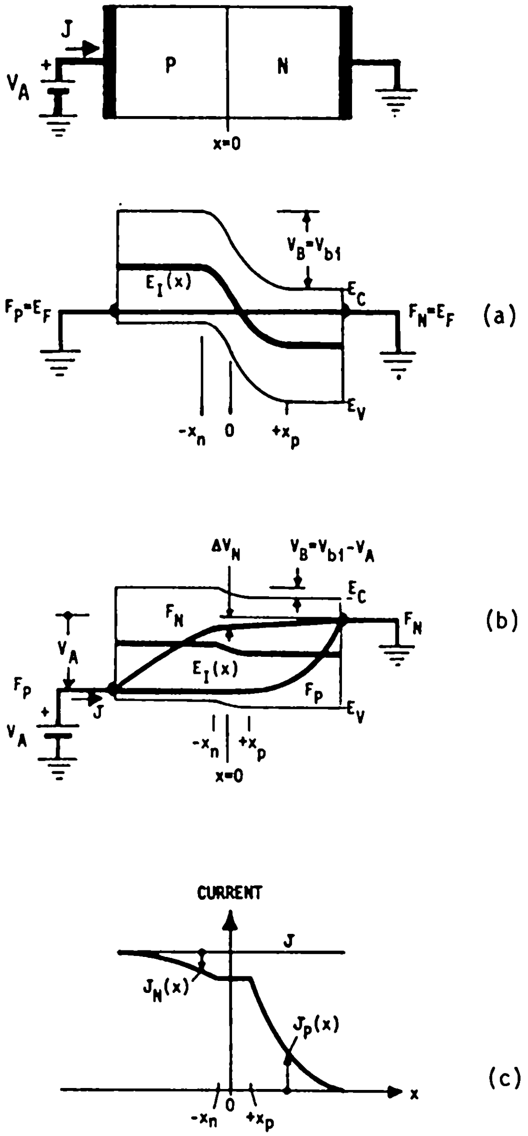

J
V_A
+
P
N
x=0
V_B=V_bi
E_I(x)
E_C
F_P=E_F
F_N=E_F
(a)
E_V
-x_n
0
+x_p
ΔV_N
V_B=V_bi-V_A
E_C
F_N
F_N
(b)
V_A
E_I(x)
F_P
F_P
+
V_A
J
E_V
-x_n
+x_p
x=0
CURRENT
J
J_N(x)
J_P(x)
(c)
x
0
-x_n
+x_p

Fig.582.1 The energy band diagrams of a p+/n junction. (a) At electrical equilibrium or zero current and (b) at a large forward bias and forward current density such that there is a significant voltage drop, ΔV_N, due to the majority carrier (electron) resistance in the space-charge layer and n-type quasi-neutral layers. (c) The current densities, $J_N(x)$, $J_P(x)$ and $J = J_N(x) + J_P(x) = \text{constant}$ versus position.

In a p+/n junction diode biased above the injection threshold voltage, V_{IT}, the terminal current is dominated by diffusion and recombination of the minority carriers (holes) which are injected into the n-type quasi-neutral layer from the space-charge layer boundary ($x=0$). It is evident that this cannot exceed the thermionic hole current flowing from the p-side to $x=-x_p$ which then passes through the space-charge layer into the n-side. In fact, the two hole currents must be exactly equal if there is no recombination in the space-charge layer. However, this hole current is affected by the voltage drop across the electron (majority carrier) resistance on the n-side. The effect of this voltage drop can be calculated using the Mott-Schottky-Davydov analysis. Refer to the energy band diagram of the p+/n junction shown in Fig.582.1(b) at a high forward bias. The voltage drop, ΔV_N, due to the majority carrier (electron) current flowing in the space-charge layer and the n-type layer can be computed using (564.2) except that J_N is no longer constant but equal to $J-J_P(x)$ in the quasi-neutral n-type layer due to the dominance of the minority carrier (hole) current in the quasi-neutral n-type layer. Let us consider a thick n-layer, then

$$J_P(x) = J_P(x=0)\exp(-x/L_p) = J_{P0}[\exp(qV_J/kT)-1]\exp(-x/L_p). \tag{582.1}$$

Thus,

$$J_N = -\mu_n N dV_N/dx = J - J_P(x) = J - J_P(0)\exp(-x/L_p) \tag{582.2}$$

which is illustrated in Fig.582.1(c). Using (564.8) for N(x),

$$N(x) = N_{DD}\exp\{(q/kT)[V_I(x)-V_N(x)]\} \tag{582.3}$$

then (582.2) can be integrated to give the modified Mott-Schottky-Davydov diode equation in which the injected minority carrier diffusion-recombination dominates. This minority-carrier-dominated Mott current is

$$J = \frac{qD_nN_{DD}[1 - \exp(-q\Delta V_N/kT)]}{\int_0^{x_{ms}+T_n} \{1 - [J_p(x_{ms})/J]\exp(-x/L_p)\}\exp[-qV_I(x)/kT]dx}. \tag{582.4}$$

It has the additional term, $[J_p(x_{ms})/J]\exp(-x/L_p)$, compared with the majority-carrier Mott diode equation given by (564.9). We shall call this new result, (582.4), the **minority-carrier Mott diode equation**. An explicit solution can be obtained if the minority carrier (electron) current in the more heavily doped p+ emitter layer can be neglected compared with the minority carrier (hole) current in the lower doped n base. Then, the terminal current from the Shockley diode equation is given by

$$J = J_p(x_{ms}) = J_{p0}\{\exp[q(V_A-\Delta V_N)/kT] - 1\} \tag{582.5}$$

which can be used to eliminate the voltage drop, ΔV_N, in (582.4) to give the general expression of the terminal current of the Shockley diode

$$J = \frac{J_{p0}[\exp(qV_A/kT) - 1]}{1 + \dfrac{J_{p0}\exp(qV_A/kT)}{qD_nN_{DD}}\displaystyle\int_0^{x_{ms}+T_n}[1 - \exp(-x/L_p)]\exp[-qV_I(x)/kT]dx} . \tag{582.6}$$

This general result reduces to the Shockley diode equation at low forward bias given by (532.12) when $J_{p0}\exp(qV_A/kT) < qD_nN_{DD}/(x_{ms}+T_n-L_p)$ for thick base or $J_{p0}\exp(qV_A/kT) < 2qD_nN_{DD}/(x_{ms}+T_n)$ for thin base. The terminal current is then

$$J \simeq J_{p0}[\exp(qV_A/kT) - 1] \tag{582.7}$$

which is identical to (532.12). At high forward bias and current when $V_A \simeq V_{bi} = (kT/q)\log_e(N_{DD}N_{AA}/n_i^2)$, the junction potential barrier is diminished to nearly zero and the terminal current becomes limited by the ohmic resistance of the n-type layer. The result is similar to that of the m/s diode given by (564.7A) and (564.7B) with J_{HB} replaced by J_{p0}

$$J \simeq J_{p0}\left[\exp\{q[V_A - J(x_{ms}+T_n-L_p)/(q\mu_nN_{DD})]/kT\} - 1\right] . \tag{582.8A}$$

This is transcendental in J for a given applied voltage V_A which can be solved for a diode current only by numerical iteration. However, just like (564.7B) for the Bethe diode, it may be inverted to give the diode terminal voltage, V_A, as an explicit function of applied current density J,

$$\begin{aligned} V_A &= (kT/q)\log_e[(J/J_{p0}) + 1] + (R_{ms} + R_{nb})J \\ &= (kT/q)\log_e[(J/J_{p0}) + 1] + \{[(T_n+x_{ms}-L_p)/(q\mu_nN_{DD})]\}J. \end{aligned} \tag{582.8B}$$

When the forward applied voltage exceeds the built-in potential, V_{bi}, (582.6) reduces to a resistance equation similar to (564.12A)

$$J = (V_A - V_{bi})[q\mu_nN_{DD}/(x_{ms}+T_n)] \tag{582.9}$$

which shows that the p+/n junction diode is represented by a built-in battery with a emf of V_{bi} volt in series with the resistance of the semiconductor space-charge layer, x_{ms}, and the resistance of the quasi-neutral layer of the lower-doped n-type semiconductor, T_n.

583 CONTACT RESISTANCE

Resistance of the metal/semiconductor contacts can limit the device current and increase the power dissipation. Hence, it is an important part of semiconductor devices. Although the Schottky rectifying barrier of the metal/semiconductor contact is highly conductive compared with a p/n junction, its conductance is still too low to make it a good (meaning low resistance) ohmic contact for the p-type and n-type semiconductors. For example, the Al/n-Si Schottky barrier has a conductance at zero applied bias of

$$\begin{aligned} G_c(V=0) &= dJ_N/dV = (q/kT)\cdot J_{HB} \\ &= (q/kT)\cdot 120(m^*/m)T^2\exp[-q(\phi_M-\chi)/kT] \\ &= (1/0.02585)\cdot 120(1)300^2\cdot\exp[-0.68/0.02585] \\ &= 4.18\text{x}10^8\cdot\exp[-0.68/0.02585]\ \ \text{S/cm}^2 \\ &= 1.58\text{x}10^{-3}\ \text{S/cm}^2 \end{aligned} \tag{583.1}$$

which is too high if used as a low resistance contact of a diode or transistor. At a current density of 1 mA/cm^2, the voltage drop across this Schottky barrier is (1 mA/cm^2)+(1.58x10^{-3} mho/cm^2) = 1/1.58 = 0.633 Volt. This is too high and would drive the Schottky barrier diode into the nonlinear forward conduction or the reverse nonconduction ranges, both of which would give high resistance.

In order to have a low resistance ohmic contact, the barrier height from the semiconductor to the metal, V_{bi} in Fig.562.1(a), must be reduced to zero as shown in Fig.583.1(a) or become attractive for the majority carriers in the semiconductor such as that shown in Fig.583.1(b). The zero semiconductor barrier height, $V_{bi}=0$, is known as the flat band condition, in analogy to that of the MOS capacitance.

Alternatively, the semiconductor can be doped to a very high donor impurity concentration at the m/s contact to make the semiconductor potential barrier very thin and to put the electron Fermi level in the semiconductor conduction band. This is shown in Fig.583.1(c). Then, the electron current density and conductance would become very high due to Fowler-Nordheim tunneling described in section 385 and the voltage drop across the contact would be small at high terminal currents.

The maximum conductance per unit area of a Schottky barrier at a forward voltage can be computed from the Bethe diode equation, (562.10). It is

$$\begin{aligned} G_c &= dJ_N/dV = (qJ_{HB}/kT)\exp(qV/kT) \\ &= (q/kT)[qm^*(kT)^2/2\pi^2\hbar^3]\exp[-q(\phi_M-\chi_S)/kT]\exp(qV/kT). \end{aligned} \tag{583.2}$$

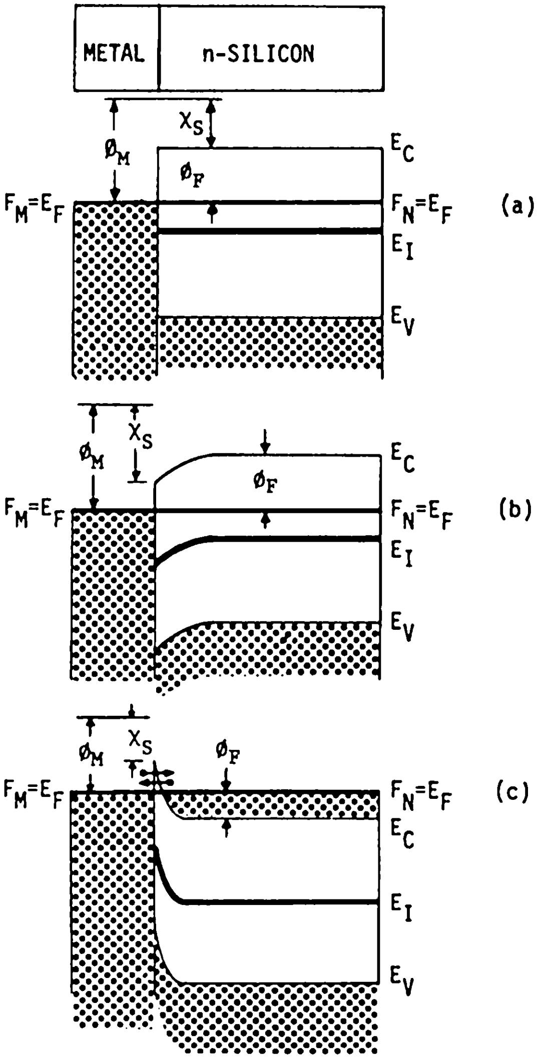


Fig.583.1 The energy band diagram of metal/n-semiconductor interfaces to give low resistance ohmic contacts. (a) The flat-band contact in which the metal-semiconductor work function difference is zero. (b) The ohmic conductance is improved further if the semiconductor work function is larger than the metal. (c) The most practical interface to give good ohmic contact is provided by tunneling when the semiconductor is heavily doped.

At zero applied voltage, V=0, (583.2) reduces to

$$G_{c0} = (q/kT)[qm^*(kT)^2/2\pi^2\hbar^3]\exp[-q(\phi_M-\chi_S)/kT) \qquad (583.3)$$

$$= (1/0.02585)\cdot 120(m^*/m)300^2(T/300)^2\exp[-q(\phi_M-\chi_S)/kT]$$

$$= 4.18\times10^8\cdot(m^*/m)(T/300)^2\cdot\exp[-q(\phi_M-\chi_S)/kT]\ \text{mho/cm}^2. \qquad (583.3A)$$

At flat-band, $\phi_M-\chi_S=\phi_F$, and this reduces to

$$G_{c0fb} = (q/kT)[qm^*(kT)^2/2\pi^2\hbar^3]\exp[-q(\phi_M-\chi_S)/kT)$$

$$= (q/kT)[qm^*(kT)^2/2\pi^2\hbar^3]\exp[-q\phi_F/kT]$$

$$= (q/kT)q(\Theta_{ave}/4)N \qquad (583.4)$$

$$= 0.02585\times4.309\times10^{-13}N$$

$$= 1.667\times10^{-11}N\ \text{mho/cm}^2. \quad (\text{or S/cm}^2) \qquad (583.4A)$$

This is the contact conductance at a semiconductor electron concentration of N that gives flat band or zero metal/semiconductor work function difference, $\phi_M-\phi_S = \phi_M-(\chi_S+\phi_F)=0$. It is identical to (kT/qJ_{TH}) and the significance of J_{TH} was noted in (562.8A). Higher donor concentration would create a repulsive semiconductor barrier for the electrons and destroy the flat band condition as illustrated in Fig.562.1(a). But it will not alter the contact conductance and resistance as indicated by (583.3) since the contact conductance depends only on the metal barrier height, $\phi_M-\chi_S$. Similarly, a lower donor concentration or electron concentration would create an attractive semiconductor barrier which would reduce the majority carrier semiconductor bulk resistance near the contact but not the contact resistance as indicated in Fig.583.1(b).

As an example, if a metal is selected such that $\phi_M=\chi_S$, then the m/s contact conductance from (583.3A) is

$$G_c = 4.18\times10^8\ \text{mho/cm}^2. \qquad (583.5)$$

If a current of 10^5A/cm^2 (the present VLSI design limit due to electromigration failure) is forced through this contact, then the voltage drop across the contact would be

$$\Delta V_c = J/G_c = 10^5/(4.18\times10^8) = 0.24\times10^{-3}\ \text{V} = 0.24\ \text{mV} \qquad (583.6)$$

This is much less than kT/q=25 mV so that the linear current-voltage approximation used in (583.6) is valid.

If we also require flat-band, although flat-band will not increase or decrease the contact resistance, then the dopant concentration in the semiconductor must be varied to make the carrier concentration equal to that given by (583.4A) or $N = N_C$. Equating (583.5) to (583.4A), then

$$G_c = 1.667\times10^{-11}N \text{ mho/cm}^2 = 4.18\times10^{8} \text{ mho/cm}^2 \qquad (583.7)$$

which gives the required electron concentration of the n-type Si,

$$N = 4.18\times10^{8}/1.667\times10^{-11} = 2.509\times10^{19} \text{ cm}^{-3}. \qquad (583.8)$$

Such a high electron concentration requires the use of the Fermi-Dirac distribution function instead of the Boltzmann approximation. This is left as a problem.

In summary, the only key to achieve lower electrical resistance in a metal/semiconductor Schottky barrier contact is to minimize the metal potential barrier height, ϕ_M-χ_S. Increase the majority carrier or dopant concentration, such as the electron in n-type phosphorus doped silicon, will not change the Schottky barrier contact resistance but will only reduce the majority carrier bulk series resistance of the semiconductor. However, at a very high concentration of majority carriers (electrons) or donor impurities, which moves the Fermi level into the semiconductor conduction band as illustrated in Fig.583.1(c), the semiconductor barrier is reduced to such a small thickness that a second majority carrier conduction path is created: the quantum mechanical tunneling through the Schottky barrier. Tunneling can provide higher current densities than thermionic emission over the Schottky barrier. When tunneling begins to dominate, the contact resistance will decrease as the dopant concentration increases. Manufacturers have invariably employed the n+ and p+ surface design to provide good ohmic contacts based on tunneling for two reasons. (i) A very high surface concentration of dopant impurity is very easy to get in the laboratory and in the factory. (ii) Small production variation of the surface concentration of the dopant impurity over the area of a six- or eight-inch silicon wafer will not increase the contact resistance significantly. (ii) is due to the fact that the surface concentration can be made so high that the tunneling current coefficient, J_{TN} in $J=J_{TN}[\exp(qV/kT)-1]$, is so many orders of magnitude larger than the operating current that slight areal nonuniformity of the surface concentration (such as 10% or even more in production environment) would not reduce J_{TN} sufficiently, even it is several orders of magnitude, to cause significant voltage drop across the ohmic contact. For example, if the tunneling current at a spot is dropped to only 10 times the operating current due to a low value of surface concentration at the spot, then the voltage drop across the spot would be $V=(kT/q)\log_e[1+(J/J_{TN})]=(kT/q)\log_e(1+0.1)=0.0953(kT/q)=2.5\text{mV}$. This would divert the current from the high resistance spot but in the worse case when the whole surface has a lower surface concentration, the total voltage drop across the contact would be only 2.5 mV which is not too detrimental to circuit operations. Evidently, the lowest contact resistance would be obtained if the metal potential barrier, ϕ_M-χ_S, is made negative to produce an attractive potential well to attract the semiconductor electrons to the metal/semiconductor interface and at the same time, the surface concentration of the dopant impurity is made very high.

590 SEMICONDUCTOR/SEMICONDUCTOR HETEROJUNCTION DIODES

Semiconductor/semiconductor (s/s) heterojunctions were proposed in the 1970's to improve the emitter injection efficiency and current gain of bipolar junction transistors using larger energy gap in the emitter than in the base. The larger energy gap greatly reduces the minority injection current in the quasi-neutral emitter layer which increases the current gain. However, there was little research effort until the 1980's when the molecular beam epitaxy (MBE) machine became available to fabricate very abrupt heterojunctions in compound semiconductors. One recent research effort focused on the Si/Ge heterojunction emitter bipolar transistor to take advantage of not only the higher emitter injection efficiency and current gain but also the higher frequency response due to higher electron mobility in the n-Ge base. Its fabrication is potentially compatible with silicon integrated circuit technology to produce monolithic chips. A second useful heterojunction is the crystalline-insulator-film/semiconductor heterojunctions (such as SiO_2/Si, SiC/Si, and Diamond/Si) to provide higher dielectric breakdown field required by submicron silicon integrated circuits with very thin gate dielectrics (< .50A) and very small device dimensions (< 1000A). A third important application is the GaAs/Si heterojunction to provide monolithic integration of GaAs light emitter and modulators with Si integrated circuits for integrated optics.

The MOS capacitor is an extreme form of a semiconductor heterojunction diode. The oxide insulator is really a large energy-gap semiconductor which does not have any electrons or holes at room temperature for conduction. Thus, the theory of MOS capacitor in chapter 4 and homogeneous p/n junctions in sections 520-55n can be combined to give the electrical characteristics.

591 Energy Band Diagram of s/s Heterojunctions

The current-voltage characteristics of s/s heterojunctions can be readily obtained from the Shockley and the SNS diode equations with appropriate modifications due to the potential discontinuities at the heterointerface boundary. The modifications are readily made using the energy band diagram which are described for ideal trap-less and real trap-ridden heterojunctions.

Trapless s/s Interface

The equilibrium energy band diagram of three possible ideal or trapless heterojunctions with $\chi_1 > \chi_2$, $\chi_1 < \chi_2$, and $\chi_1 << \chi_2$, are shown in Figs.591.1(a), (b), and (c). Appropriate impurity doping can be selected to give the flat-band energy band diagrams shown in Figs.591.1(d), (e) and (f). The construction procedure is identical to that of the MOSC given in Figs.413.1(a)-(d), the p/n junction in Figs.521.1(a)-(e), and the s/m Schottky barrier in Figs.561.1(a)-(c). Energy levels are not labeled in Figs.591.1(d)-(f) but are self-evident.

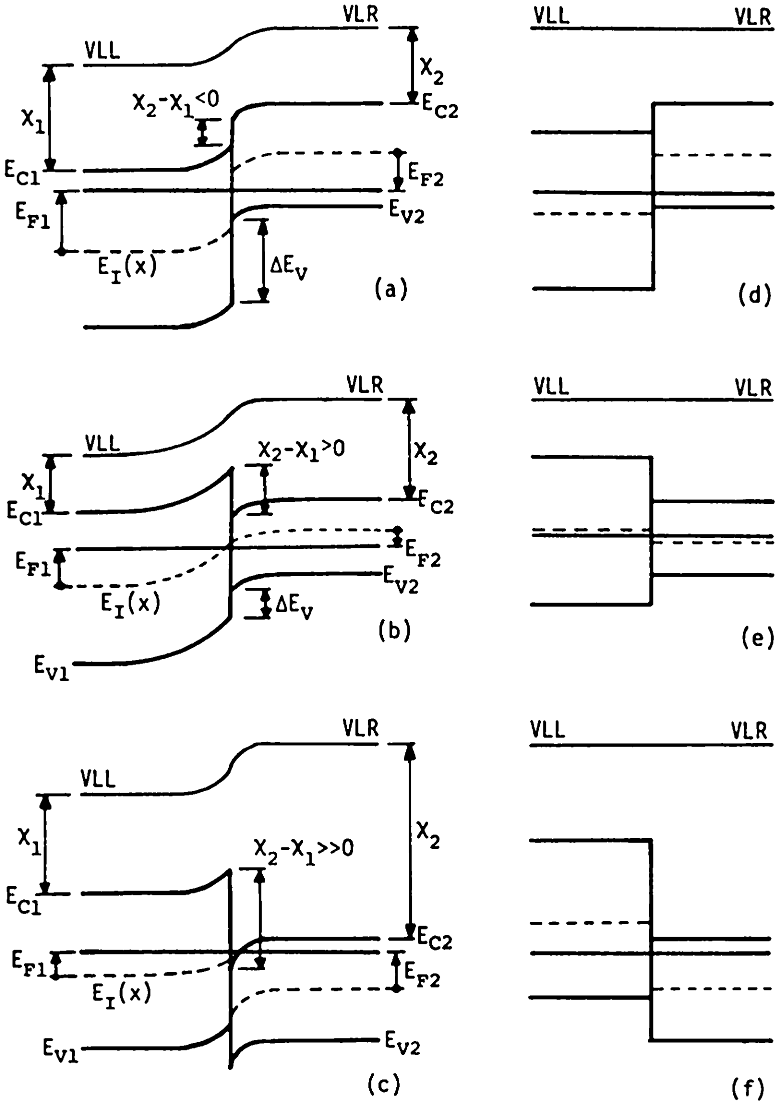

VLR
VLL
X_1
X_2
$X_2-X_1<0$
E_{C1}
E_{C2}
E_{F1}
E_{F2}
E_{V2}
$E_I(x)$
ΔE_V
(a)
VLL
VLR
(d)
VLR
VLL
X_1
X_2
$X_2-X_1>0$
E_{C1}
E_{C2}
E_{F1}
E_{F2}
E_{V2}
$E_I(x)$
ΔE_V
E_{V1}
(b)
VLL
VLR
(e)
VLR
VLL
X_1
X_2
$X_2-X_1>>0$
E_{C1}
E_{C2}
E_{F1}
E_{F2}
$E_I(x)$
E_{V1}
E_{V2}
(c)
VLL
VLR
(f)

Fig.591.1 Energy band diagrams of semiconductor/semiconductor heterojunctions with no interface traps. (a) $X_2<X_1$. (b) $X_2>X_1$. (c) $X_2>>X_1$. (d),(e), and (f) are corresponding cases at flat band.

A unique feature of the s/s heterojunction is the discontinuity of the conduction and valence band edges at the heterojunction boundary. These are labeled in Figs.591.1(a) and (b). They are defined in figure (b) (with sign) by

$$\Delta E_{C21} = (X_2 - X_1) \tag{591.1A}$$

$$\Delta E_{V21} = E_{G2} - E_{G1} - (X_2 - X_1) \tag{591.1B}$$

and

$$\Delta E_{C21} + \Delta E_{V21} = E_{G2} - E_{G1} \tag{591.2}$$

These are fundamental properties of the material and hence are invariant. In practice they vary due to the presence of interface traps described later.

The amount of energy band bending in the two semiconductors depends on the impurity doping concentration. Figures 591.2(a)-(c) show the flat band condition produced by adjusting the doping impurity concentration on the two sides of the semiconductor. Fig.591.2(b) shows obviously that flat band is attained when

$$(X_2 - X_1) = (E_{C1}-E_F) - (E_{C2}-E_F) \tag{591.3}$$

$$= kT\cdot\log_e(N_{B1}P_{B2}/n_{i1}n_{i2}) \tag{591.3A}$$

The Boltzmann approximation for the carrier concentration is used in (591.3) to give (591.3A). It is obvious from the energy band diagrams shown in Figs.591.1(d)-(e) that the flat band condition can be attained with an infinite number of N_{B1} and P_{B2} or N_{DD1} and N_{AA2} combinations which is also indicated in (591.3A).

Trappy s/s Interface

Semiconductor/semiconductor interfaces are not ideal with perfect bonding even if it is made in ultra-clean environment of high vacuum so that no interfacial layer is formed due to foreign atoms or contaminations. They are not ideal or perfect because they contain a significant number of dangling bonds due to the different lattice constant or interatomic spacing of the two semiconductors and due to thermal differential cooling which produces tremendous strain by the lattice mismatch. The dangling bonds are interfacial traps. If charged, the traps will alter the amount of energy band bending at the two semiconductor surfaces. This bending will give an apparent s/s electron barrier height which differs from the work function difference, ΔE_{C21}, given by (591.1A). The interfacial layer is fabrication process dependent and hence random variation of the apparent work function deduced from the current-voltage characteristics are frequently observed. One can illustrate this by assuming that there is an interfacial SiO_2 of 2A thick on the Si surface causing $10^{13}cm^{-2}$ positively charged dangling-bond traps. Then the electron barrier height or conduction band discontinuity is given by

$$\Delta E_{C21}/q = (X_2 - X_1)/q - qN_{OT}x_o/\varepsilon_d \tag{591.4}$$

Using these numbers, the electron energy barrier will be reduced by $qN_{OT}x_o/\varepsilon_d$ = $1.6\text{x}10^{-19}\text{x}10^{13}\text{x}2\text{x}10^{-8}/(4\text{x}8.85\text{x}10^{-14})$ = 0.1Volt. This barrier lowering will increase the Shockley diode current 55 folds, exp(0.1/0.025)=exp(4)≈55.

Lowering of the barrier height is illustrated in Fig.591.2(a) by a sheet of positive charge trapped in the interfacial oxide in a n-Si/SiO$_2$/p$^+$Ge diode. Comparing with the trapless and oxideless n-Si/p$^+$Ge diode shown in Fig.591.2(b), it is evident that the electron barrier is significantly reduced (pulled downwards) by the positive oxide charge. The diode current is dominated and limited by minority carrier (electron) diffusion into the p$^+$Ge from the interface if the oxide barrier is so thin that the electron tunneling rate is very high.

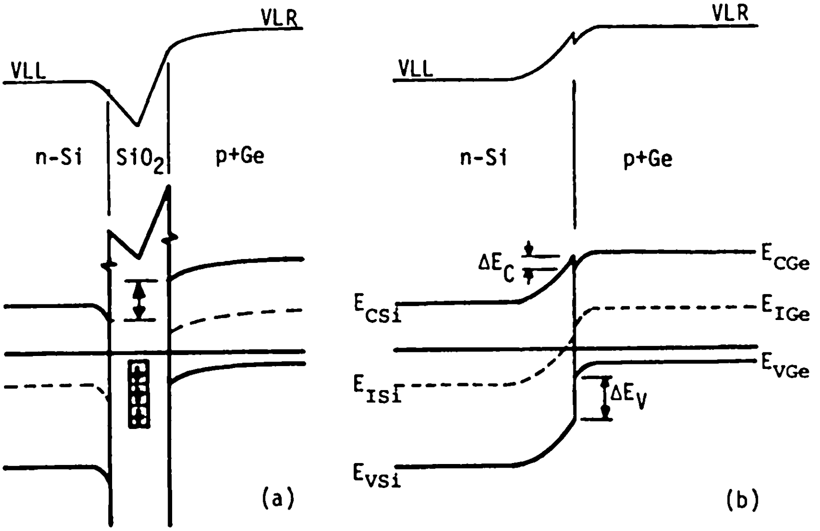


Fig.591.2 Effect of interfacial traps on the energy band of a heterojunction diode. (a) Trappy p-Si/SiO$_2$/n-Ge diode. (b) Trapless p-Si/n-Ge diode.

592 Electrical Characteristics of s/s Heterojunctions

The mathematics to derive the dc, ac, and transient characteristics of a s/s heterojunction are identical to those for the p/n junction described in sections 53n-55n. The only change is that there are two intrinsic carrier densities, n_{i1} and n_{i2}, and that the built-in potential or the work function difference contains the electron affinity difference which is no longer zero. Thus, the Shockley and SNS diode equations can be modified to give the d.c. I-V characteristics. The modifications can be made by inspection of the trapless and trappy energy band diagrams given in Figs.591.1 and 591.2.

599 BIBLIOGRAPHY

There are many solid-state and semiconductor device books with sections and chapters on the electrical properties of p/n junctions. Some books also covers the theories of the metal/semiconductor Schottky barrier diodes and the semiconductor/semiconductor heterojunction diodes. Few discuss the contact resistance. None covers the capacitance and current transients described in section 554. This bibliography gives a selected chronological lists of textbooks and research-engineering monograms on p/n, m/s, and s/s junctions of historical significance and detailed analyses. A short description of each book is given to guide the students for further reading. These books were also used by the author to learn the history of the semiconductor diode theory and to teach advanced undergraduate and beginning graduate device courses for many years.

[599.1] R. H. Fowler (Cambridge University), **Statistical Mechanics**, 2nd ed., section 11.9, pp.429-436, Cambridge University Press, 1936. Paperback edition 1966. This is the first book containing a sections on contact rectifiers or solid/solid junctions. The rectification (I-V) equations are derived for a metal/insulator/metal diode due to thermionic emission over the barrier (Bethe diode) and tunneling (like the Easki p+/n+ diode). The treatment is advanced but the result is familiar when compared with those of sections 56n. It was used in the following reference [599.2] and evidently also used to derive the tunnel I-V characteristics of p+/n+ diodes when Esaki discovered and explained the forward negative resistance in 1958.

[599.2] Henry C. Torrey and Charles A. Whitmer (Rutgers University), **Crystal Rectifiers**, MIT Radiation Laboratory Series Vol.15, McGraw-Hill Book Co. 1946, reprinted by Boston Technical Publishers, Inc., 1964. This is the first book on semiconductor rectifier diodes using the metal(tungsten whisker)/semiconductor(Si) Schottky barrier. It gives the historical record of the U.S. World-War-II effort on radar detector development, including the various I-V theories in chapter 4 leading to final successful theory of Bethe. This historical volume has been known as 'Torrey-Whitmer' to the pioneers of solid state devices in the late 1940's and the 1950's.

[599.3] Robert V. Pound (Harvard), **Microwave Mixer**, MIT Radiation Laboratory Series Vol.16, McGraw-Hill Book Co. 1964, reprinted by Boston Technical Publishers, Inc., 1964. This book contains the microwave characteristics and applications as a radar detector.

[599.4] William Shockley (Bell Labs., Shockley Transistor Corp., Stanford), **Electrons and Holes in Semiconductors**, D. Van Norstrand Co. Inc. 1950. This is the first book containing the modern theory of the Shockley (p/n junction) diode. The description and derivation are covered in several chapters (sections 4.2, 4.3, 12.5) and a short history starts on page 3 in chapter 1.

[599.5] R.D. Middlebrook (Stanford, Caltech), **An Introduction to Junction Transistor Theory**, John Wiley & Sons, Inc. New York, 296pp, 1957. This is first book devoted entirely to p/n junction diode and transistor characteristics. Chapters 5, 6 and 7 covers the p/n junction. It was a reprint of the author's doctoral thesis in EE at Stanford. Particularly clear pictures of the carrier concentrations in the quasi-neutral layers are given throughout the three chapters.

[599.6] Eberhard Spenke (Pretzfeld), **Electronic Semiconductors**, McGraw-Hill Book Co. Inc., 402pp, 1958. This is the first and perhaps the only book that gives a very detailed treatment of the p/n and m/s diodes with carefully defined energy band diagram (chapter X) like our chapter. Unfortunately the notation choice was not optimum making it unnecessarily hard. An elementary treatment without solving any differential equations is given in chapter IV. Like the previous books, recombination-generation currents from traps in the junction space-charge layer are neglected.

[599.7] A.K. Jonscher (GE Ltd, Wembley), **Principles of Semiconductor Device Operation**, John Wiley & Sons, 168pp, 1960. This is the first book that included p/n junction characteristics beyond the Shockley diode theory. Chapter 4 contains the SNS space-charge layer generation-recombination current, conductivity modulation and space-charge-limited currents at high forward injection levels, and interband impact generation current at high reverse bias. Because of this extended coverage, it was the popular textbook for serious, nonpedstrain introductory device courses taught during the 1960's. Some high-injection-level results of the transistor sections were erroneously used by later textbook authors as well as researchers when Jonscher's simplifying assumptions were overlooked.

[599.8] Alvin B. Phillips (Motorola), **Transistor Engineering**, McGraw-Hill Book Co. New York, 373pp, 1962. This is the first of three books (see below) that gave extensive analysis of the d.c.,

small-signal, and transient characteristics of p/n junctions to form a base for the comprehensive analysis of the two-junction bipolar transistors. They are given in chapters 5 and 6. High level results have restricted utility due to limitations in the assumptions.

[599.9] Paul E. Gray (MIT), David DeWitt (IBM), A.R. Broothroyd (Queen's College), and James F. Gibbons (Stanford), **Physical Electronics and Circuit Models of Transistors**, John Wiley & Sons, New York, 262pp, 1964. This is the second volume of a seven-volume textbook series especially written for junior electrical engineering students by the SEEC (Semiconductor Electronics Education Committee) supported by the U.S. National Science Foundation beginning in fall-1960, composed of veteran college teacher-researchers and industrial researchers. The manuscript was revised many times based on teaching the materials at both universities and industrial laboratories during 1960-1964. It is still the most concise, accurate and eloquently presented textbook on p/n junction diodes and transistors. It is also the first book that contains the circuit model of the device.

[599.10] Joseph Lindmayer and Charley Y. Wrigley (Sprague Electric Co.), **Fundamentals of Semiconductor Devices**, D. Van Nostrand Co. Inc. New York, 485pp, 1965. This is the first book that gives detailed analysis of not only of p/n (chapter 2) but also p/i diodes (chapter 7). Correct transient solutions for switching transients are given.

[599.11] S.M. Sze (Bell Labs), **Physics of Semiconductor Devices**, John Wiley & Sons, 812pp, 1969, 868pp, 1981. Chapters 2 and 5. This book has been the standard reference on devices since its publication.

[599.12] A.G. Milnes (CMU) and D.L. Feucht (CMU, SERI), **Heterojunctions and Metal-Semiconductor Junctions**, Academic Press, 408pp, 1972. This and the following three books focus on heterojunctions and are the only three books entirely devoted to heterojunctions, but the MOS heterojunction is not included.

[599.13] B.L. Sharma and R.K. Purohit (SSPL, Delhi), **Semiconductor Heterojunctions**, Pergamon Press, 216pp, 1974. This book also include the energy band properties of many compound semiconductor heterojunctions. It has a tabulation of the I-V formulae of all the possible energy band offsets at the heterojunction boundary.

[599.14] E.H. Rhoderick (Manchester), **Metal-Semiconductor Contacts**, Clarendon Press, Oxford, 201pp, 1978. This gave probably the most detailed treatment of the theory of rectifying Schottky barrier diode. However, compare the results with those given in sections 56n and 571.

[599.15] Arthur B. Glaser (BTL) and Gerald E. Subak-Sharpe (CCNY), **Integrated Circuit Engineering**, sections 2.1 to 2.9, pp.15-35, Addison-Wesley Publishing Co., 1979. This book gives many design charts to compute the circuit elements of p/n junction diodes.

599 PROBLEMS

P512.1 Calculate the exact dimensions of the cross section of the interstitial channel in the <110> and <100> directions in the diamond and zinc blend lattices in terms of the lattice constant a. Obtain the numerical values in angstrom unit (10^{-8}cm) for Si. (Use the unit cell figures and lattice constants given in chapter 1.)

P512.2 The temperature dependence of E_G in (512.2) used to compute its value in (512.3) at the diffusion temperature of 1300K gave an apparent activation energy for substitutional diffusion of $7E_G \cdot 4$eV consistent with data. Show that if this temperature dependence is plugged into (512.1), the theoretical thermal activation energy would be too high. Is there a discrepancy or inadequacy in this analysis and how could it be improved if there is?

P512.3 An alternative to the atomic model used to estimate the activation energy for diffusion is to use the thermodynamic functions and their experimental values as well as the experimental chemical bond energies. However, these values are usually given at the standard condition, 298K and 1 atmosphere, with the solid as the reference state. Nevertheless, many attempts have been made using this approach. Given the following, come up with an estimated thermal activation of substitutional diffusion. Si-Si(gas) = 78.1±2.4kcal/mole = 136.4±12.6kJ/mole = 1.4137eV. ΔH^{o}(298K) = 108±2kcal/mole = 450±8kJ/mole = 4.664eV. Heat of formation of gas atoms in their standard state $G = H - TS$, $\Delta G = \Delta H - S\Delta T$. ΔG^{o} = 323.9kJ/mole = 3.357eV. ΔH^{o} = 368.4kJ/mole = 3.818eV. S^{o}(g) = 167.86J/mole-K = 1.7397meV/K and S^{o} = 18.7J/mole-K = 0.194meV/K.

P512.4 First generation (1960) ultrahigh speed switching Si diodes, and transistors such as the 1N914 and 2N706 used the the first supercomputer, the Control Data Corporation model 6600, obtained their speed by doping with the gold recombination center to a concentration of 10^{16}cm^{-3} because gold has two very efficient recombination energy levels (see Fig.381.2) and gold can be diffused into Si quickly at low diffusion temperatures due to its high diffusivity in Si (see Fig.512.2). Suppose that this high-temperature gold diffusivity in Si can be extrapolated to room temperature (300K) using the Arrhenius relationship. (a) What is the gold diffusivity in Si at 300K? and (b) what is the time to failure because the substitutional gold atoms jump (a diffusion step) to adjacent interstitial sites 5A away and interstitial gold is not an electronic trap.

P512.5 It has been generally known to Si technologists (such as the CCD or Charge Coupled Device and high efficiency solar cell engineers) that iron impurity is a rather nuisance recombination center because it is highly unstable and its recombination activity disappears in a few days after the device is fabricated and sitting on the shelf. (a) Demonstrate the fast diffusivity of Fe in Si by extrapolating the high-temperature diffusivity of Fe in Si (Fig.512.2) to room temperature. (b) If the generation-recombination volume of the CCD or solar cell is in a thin Si surface layer (about 1000A to allow efficient light penetration), and there is a sheet of atomic sink for Fe at 10μm from the Si surface, what is the duration of this instability?

P512.6 SiO_2 clusters are found to be excellent sink for metallic impurities in Si crystals. Experiments have shown that SiO_2 clusters can be formed inside Si owing to oxygen precipitation at defect (dangling bond) sites during prolonged heating of Si if the Si is supersaturated with oxygen at the heating temperature. Supersaturation means that the oxygen concentration is higher than the maximum solubility at the observation temperature. Suppose that an SiO_2 cluster is formed at 1 μm below the Si surface. How long does the SiO_2 cluster disappear into the Si surface at room (300K) and how long is the maximum shelf storage (125C) temperatures of an Si diode or transistor? Use Fig.512.2 to extrapolate or use the equation in the figure to calculate the oxygen diffusivity in Si at 300K and 125C.

P513.1 A p+/n junction is to be fabricated by a two-step diffusion procedure involving an initial pre-deposition step of an amount, Q (atom/cm^2), of boron impurity atoms and a final drive-in step at a higher temperature to drive in the deposited impurity atoms by diffusion to a required junction depth, x_j (μm), and to achieve a required surface concentration of C_0 (atom/cm^3). The substrate n-type donor dopant impurity concentration is C_B (atom/cm^3). Given: $C_0=1.0x10^{19}cm^{-3}$, $C_B=1.0x10^{15}cm^{-3}$, $x_j=1.0\mu m$, and $T_d=1000°C$, (a) What is the diffusion time required? (b) What is the predeposited Q?

P522.1 A diffused Si p+/n junction is to be approximated or modeled by an abrupt Si p+/n junction. The abrupt model is defined by $N_{AA}=1.0x10^{19}cm^{-3}=$constant from $x=0$ to x_j and $N_{DD}=1.0x10^{15}cm^{-3}=$constant from x_j to the back surface and $x_j=1.0\mu m$. What are the following properties at equilibrium, $V_{Applied}=0$, at $T=296.6K$ at which $n_i\approx 1.0x10^{10}cm^{-3}$ and $kT/q\approx 0.025$ mV? (a) The potential barrier height, V_B, (which is V_{bi} at $V=0V$) in (mV). (b) The total space charge layer thickness, x_{pn} (μm). (c) The space charge layer thicknesses on the p-side and n-side, x_p and x_n in (μm). (d) The maximum electric field, E_{MAX} (V/cm).

P522.2 Derive the V_{bi} formulae using $J_P=0$ and the procedure like that in the text for $J_N=0$.

P522.3 The equilibrium barrier height expression (522.5) is very general. Show that it is valid for an exponentially graded p/n junction where N_{DD}-N_{AA} = $C(x)$ = $C_B\{1\text{-}\exp[-(x\text{-}x_j)]\}$ where x_j is the junction depth, C_B is the bulk concentration of the donor and the surface is at $x=0$ and the surface concentration is $-C(0)$ = $C_B\{1 + \exp(x_j/L)\}$ where $x_1=x_p$ and $x_2=x_n$ are the two boundaries of the space-charge layer at equilibrium.

P522.4 The built-in potential formulae given by (522.7A) made use of three approximations: extrinsic, complete ionization, and Boltzmann assumptions. Generalize this by removing one, two or all three approximations with the exact nonextrinsic solution given by (242.11)-(242.14A), and the deionization and degenerate (Fermi) solutions in sections 25n. (a) Nonextrinsic but still Boltzmann approximation. Your solution can also be expressed in terms of hyperbolic functions. (b) Extrinsic, Boltzmann, but includes deionization. (c) Other permutations of deionization and degeneracy in the extrinsic range.

P522.5 Diodes and transistors are no longer useful when the temperature is so high that the p/n junction built-in potential barrier height is less than about 4kT. Calculate this temperature for an Si p+/n abrupt junction diode with $N_{AA}=10^{18}cm^{-3}$ and $N_{DD}=10^{16}cm^{-3}$.

P523.1 Sketch to scale five diagrams like those of Figs.523.1(a)-(e) at equilibrium for the diode in problem P522.1 using the numerical results obtained in problem P522.1. The n-type VLSI Si wafer has a diameter of 8 inches and is 500μm thick. In order to clearly illustrate the spatial variations along the thickness or depth direction, plot only a small and expanded surface layer from the Si surface at $x=0$ to $x=n\cdot x_j$. You are to select a value for 'n' (about 1.1 to 2.0 times x_n would be sufficient) based on the x_p and x_n values you computed in problem P522.1 so that the $V_I(x)$ variations on both sides of the p/n junction boundary are clearly illustrated without exceeding the width of your 8.5" wide paper. If one side still cannot be shown clearly, give also an expanded plot of the x-axis on that side. Note: different origins of the x-axis are used, one for the junction depth x_j and a second to compute x_n and x_p as well as the potential drops on each side. The different origins are selected so that the analytical solutions have the simplest algebraic expression to clearly reveal the physics and the important parameters at one glance.

P531.1 The p+/n junction in problem P522.1 is now biased. (a) Calculate the parameters asked for in problem 5.2 at $V=+0.5V$, and $V=-10V$. (b) Display your results in a table with three columns, one each for $V=-10V$, 0V (from answers in problem P522.1) and +0.5V. (c) Sketch to scale the energy band diagram, the space-charge density including P(x) and N(x) in semi-log scale, and E(x) at these two biases, using Figs.531.1 and 531.2 as guides.

P534.1 A forward voltage of V is applied to a thin-base p+/n/m diode whose n-base is 1μm thick, i.e. $T_n=1\mu m$. And $T_n<<L_p=\sqrt{D_p\tau_p}=100\mu m$. The n/m back surface contact is perfect (zero resistance). Sketch P(x) and $J_P(x)$ in the n-type quasi-neutral base layer without solving the hole diffusion equation. Label the concentrations and current densities at the boundary x_n and T_n.

P534.2 An Si p+/n junction is fabricated on a graded n-type Si with a donor in-diffusion profile or epitaxial profile given by $N_{DD}(x)=G[(x/x_1)+1]^{-1/2}$ where $G=10^{16}cm^{-3}$, $x_1=1\mu m$. The p+ layer has $N_{AA}=\text{constant}=10^{19}cm^{-3}$. Using the depletion approximation, find the spatial variation of (a) the electric field, and (b) electric potential at V_A; and (c) the value of the maximum electric at $V_A=0$. (Answer: $E(x)=E(0)+(qG/\varepsilon)2x_1\{\sqrt{[1+(x/x_1)]}-1\}=-2(2G/\varepsilon)x_1\{\sqrt{(1+y_n)}-\sqrt{(1+y)}\}$ where $y_n=x_n/x_1$. $V_{bi}=(2\varepsilon G/3q)x_1{}^2\{2-(2-y_n)\sqrt{(1+y_n)}\}=0.889V$ and $E(0)=5.09\times10^4V/cm$.)

P536.1 An Si p+/i/n+ diode has a nearly pure ($N_{AA}=N_{DD}=0$) 10μm i-layer. (a) What is the breakdown voltage assuming $\alpha=(\alpha_n+\alpha_p)/2$ and using Fig.384.1. (b) What is the electric field at breakdown? (c) What is the electric field at 10% of breakdown? (d) Give at least three important fundamental material parameters that can affect the temperature variation of the breakdown voltage. (e) Which of the factors dominate the temperature dependence?

P537.1 Continue on to solve for the remaining characteristics of the p+/n diode given in problems P522.1, P522.5 and P523.1. Assume $\tau_p=\tau_n=1.0\mu s$, $n_i=1.0\times10^{10}cm^{-3}$, and $kT/q=0.025V$ (T=296.6K). Use the appropriate majority and minority carrier mobilities from figures or formulae in chapter 3. (a) Calculate J_1 of the Shockley and J_2 of the SNS current components in A/cm^2. (b) Look up the breakdown voltage of the junction from Fig.536.2. (c) Compute the threshold injection voltage as defined by $J_{SHOCKLEY}=J_{SNS}$ discussed at the end of section 535. This will help to plot the following figures. (d) Sketch the I-V curves in the three I and V scales similar to those used in Fig.500.1. Select the appropriate range of I and V scale for each of the three figures in order to show the details of the variations. Read hints below to do the sketches in order to save time while still get very accurate plots. (**Hints for the sketches:** You do not need and are not asked to compute many points of the I-V curves in order to plot the three figures. If you do, you will have spent lots of time doing a technician's job instead of understanding the basic physics. What you need to do is to make use of the $\log_e I$ behaviors and the J_1 and J_2 values you just computed to draw the **two** straight line segments (from Shockley and SNS current components) in the forward quadrant of Fig.500.1(c), $\log_e I$ versus V. For the reverse quadrant, using one more I-V point at a reverse bias, such as the $V = -10V$, whose x_{pn} you just computed in problem 531.1, you can also draw **accurately** the **three** straight line **segments** (ohmic; $I\propto x_{pn}\propto\sqrt{(V_{bi+}V_R)}$; and BV) in the reverse quadrant, $\log_e I$ versus $\log_e V$. The segments can then be combined to give the total current. The linear-I versus linear-V curves of the first two figures can then be sketched by reading off a few points from the expanded I-V curves just obtained in the logI-V and logI-logV plots. Thus, you need a semilog paper and a log-log graph paper such as the ones used in Fig.500.1(c) to speed up doing this problem. The instructor may provide you these papers or you may make a copy of Fig.500.1 and plot your lines and curves on them with color pencils.)

P537.2 Derive the theoretical formulae for the diode temperature at which the Shockley and SNS current components are equal.

P537.3 Derive the theoretical formulae for the lifetime at which the Shockley and SNS current components are equal.

P537.4 Derive the theoretical formulae for the dopant impurity concentration of the lower-doped side of a p++/n or n++/p junction at which the Shockley and SNS current components are equal.

P537.5 Given $N_{AA}=10^{19}$ and $N_{DD}=10^{17}\ \text{cm}^{-3}$, $\tau_{n0}=\tau_{p0}=10^{-9}\text{s}$, and T=22C, what is the threshold injection voltage of an Si diode? Use the Si material data given in the figures of chapters 2 and 3.

P540.1 If the charge storage on the traps in the space-charge layer cannot be neglected, where is the capacitor, C_t, connected in the circuit model shown in Fig.540.1(b).

P541.1 Verify the d.c. electron charge stored in the quasi-neutral p-layer, Q_N given by (541.7), using the same algebra as those leading to Q_P given by (541.5).

P541.2 We used the absolute sign for Q_D in C_d of (541.11) to focus on the physics. Find out the sign from the space charge versus position diagram of Figs.531.1(a) and (b) if $C_d=dQ_D/dV$ and $Q_D=Q_P$. What if $Q_D=Q_N$?. Give either algebraic proof or physical reason or both.

P541.3 A common question asked by the beginner first exposed to device theory and physics is 'Why are C_n and C_p not in series with a parallel combination of C_{pn} and C_d since they come from three series layers?' Similarly, 'Why are G_n, G_p and G_{pn} not in series but are in parallel?' Give the physical reason. The mathematics is already given in the text.

P541.4 What is the significance and physics error when $V=V_{bi}$ which makes $x_{pn}=0$ and $C_d=\infty$?

P541.5 Show mathematically that the majority carrier storage capacitances in the two QNL's of the p/n junction are already included in C_p and C_n.

P542.1 Verify the numbers computed in Table 542.1.

P542.2 Calculate a table like Table 542.1 at V=-5.0V, V=0.025V, and V=0.23V. Verify some of the values at V=0.23V with your expectation when V is equal to the injection threshold.

P542.3 A silicon p/n junction diode is used as a voltage tunable capacitor to automatically scan the AM radio stations in the 550kHz to 1650kHz band. What is the applied d.c. voltage range needed to tune this 3:1 frequency range assuming that you have a 12 Volt battery? (A tuned parallel L-C circuit has a resonance or peak impedance frequency of $f=1/2\pi\sqrt{LC}$.)

P542.4 What is the substrate dopant concentration needed to make a p++/n junction capacitor to tune over the 3:1 frequency range of the 550kHz to 1650kHz AM radio band with a variable voltage of 2-12 Volt. The diode is to have a diameter of 10 mils or 250 micrometers and the inductance of the tuned LC circuit is 100 microhenry? (A tuned parallel L-C circuit has a resonance or peak impedance frequency of $f=1/2\pi\sqrt{LC}$.)

P542.5 A long (or thick) silicon p/n junction diode, such as the numerical example given in the text, is made on a silicon wafer of 1 mm thick so that the series bulk resistances cannot be neglected. Calculate the cutoff frequency due to the series resistance when the diode is reverse biased, zero, and forward biased at the three voltages in Table 542.1. How does these cut-off frequencies compare with the diffusion-recombination delay (the diffusion-recombination delay time is roughly the lifetime or the G/C time constant and the charging time of the space-charge layer)? Under what application conditions is the series resistance important?

P542.6 Sketch to scale on one linear graph paper the reverse C-V curve of the following three Si p/n junction diodes all with an area of $100\mu\text{m}^2$. Plot C from 1 to 31 fF (use 4 fF/inch) and V from 0 to -20V (use 2 V/inch). You need not computer many points of (C,V) but give your reasoning concisely. (a) m/p+/n/m with $N_{AA}=10_{19}\text{cm}^{-3}$ and $N_{DD}=10^{16}\text{cm}^{-3}$. (b) m/p+/n/n+/m with $N_{AA}=10^{19}\text{cm}^{-3}$, $N_{DD}=10^{16}\text{cm}^{-3}$ and a thickness of 0.5μm, and $N_{DD}{}^{+}=10^{19}\text{cm}^{-3}$. (c) m/p+/n/n-/n+/m with the same doping concentration and n-layer thickness, and a 9.5μm n- layer of $N_{DD}{}^{-}=10^{14}\text{cm}^{-3}$.

P552.1 The turn-on transient analyzed in section 552 was for the voltage response due to an applied current step. (a) Why didn't we find the current response due to an applied voltage step? (Hint: $i = CdV/dt$.) (b) Sketch the current waveform if a forward voltage step of V_A is applied through a resistance whose value is $R = (V_A - V_{bi})/J_F$.

P553.1 Let the forward current be 1 mA and the reverse voltage be 5V applied through a series resistance of 50 ohms. Let the minority carrier lifetime (holes) in the p+/n diode be 1 microsecond. What are the turn-off time constants? Does the expectation $t_I < t_S < t_I + t_{II}$ still hold? Give reasons if it does and if it does not. State also which is a better quantity to use in a practical application (t_S if $t_S >> t_I + t_{II}$) and why?

P553.2 Let the forward current be 1000 mA or 1 ampere and the reverse voltage be 5V applied through a series resistance of 50 ohms. Let the minority carrier lifetime (holes) in the p+/n diode be 1 microsecond. What are the turn-off time constants? Does the expectation $t_I < t_S < t_I + t_{II}$ still hold? Give the reasons if it does and if it does not. Compare your results with the example in the text. Discuss the ways you would design your diode and specify the applied voltages and currents if you are asked to produce a very fast reverse recovery diode, and if you are asked to produce a diode that will produce a specified long delay before the switch is opened.

P553.3 Give a simple reason based on device physics or charge store concept to show that the speed up of the recovery during switch-off transient is given by the ratio, J_F/J_R. Afterwards, show this also mathematically using the solutions given in the text.

P553.4 Obtain the switching transient solutions for a thin-base diode with the p+/n/m structure where n is the thin base of geometrical thickness X_B and $X_B << L_B$. n/m is a perfect ohmic contact of infinite recombination rate. Use the charge control model and assume minority carrier diffusion dominant. The mathematics is very simple.

P553.5 Obtain the switching transient solutions for a thin-base diode with the p+/n/n+ structure where n is the thin base of geometrical thickness X_B and $X_B << L_B$ but recombination in the thin base-layer must not be omitted. n/n+ is a perfect donor contact of zero minority carrier (electron) current. Use the charge control model and assume minority carrier diffusion dominant. The mathematics is very simple.

P554.1 Design a n+/p abrupt Si junction diode to measure the emission and capture rates of electrons and holes at the gold donor level $E_V + 0.35\text{eV}$ in the Si energy gap. Let $N_{DD}(n+) = 10^{18}\text{cm}^{-3}$, $N_{AA} = 10^{15}\text{cm}^{-3}$, $N_{TT}(\text{Au}) = 10^{13}\text{cm}^{-3}$, and diameter of the junction be 30mils or 30x25.4μm. How sensitive must the capacitance and current meters be to measure the transients at 400K, 300K, and 200K using a reverse voltage of 10V?

P554.2 A n+/p heterojunction diode of the same donor and acceptor dopant concentrations and areas as the Si homojunction diode above is known to have a layer of interfacial traps located at the metallurgical boundary. What is lowest density (trap/cm^2) detectable using the current transient and the capacitance transient? Assuming a capacitance noise of 0.1fF and a current noise of 0.1fA. At what temperatures would you use these two methods to get maximum sensitivity and sufficiently slow decay to allow the use of highly sensitive but slow capacitance and current meters?

P561.1 Practice the steps of drawing the equilibrium energy band diagram and charge density distribution of an Au/p-Si metal/semiconductor diode following Figs.561.1(a)-(d). Give your figure with energy levels to scale for Au and p-Si with $N_{AA} = 10^{17}\text{cm}_{-3}$.

P562.1 The hot electron current from the metal to the semiconductor in a m/s junction, (562.9A), was not explicitly derived but was obtained using the zero current condition at equilibrium or zero applied voltage. Derived J_{Nx}(metal→semiconductor) by carrying out the integration on the metal side of the m/s boundary using a procedure similar to (562.7) for Si.

P563.1 If the Al/n-Si Schottky barrier contact shown in Fig.563.1 was used as the ohmic contact to the n-Si surface of the 1957SNS p+/n diode, how would the forward I-V characteristics look? Sketch it both on linear-linear and semi-log paper by reading off a few important points from the two forward curves in Fig.563.1. Are the reverse characteristics affected? What can you conclude from this exercise - what can an experimental I-V curve reveal about the processing and design difficulties?

P564.1 Assuming complete ionization and $N_{DD} = \text{constant} \neq f(x)$ and depletion for the Schottky barrier, find the formulae of the Bethe diode current that includes the Mott correction. Use

(564.10) and the depletion approximation of the space-charge layer previously obtained for the p++/n diode. Does this solution, which is based on the depletion approximation, give the correct asymptotic result at high forward voltage when $V_A \rightarrow V_{bi}$ and why?

P570.1 What is the peak current density and peak voltage for a abrupt p+/n+ tunnel diode in a direct material? What is the valley current if N_{TT} traps are in the space-charge and quasi-neutral layers? (Hint: Results are in terms of the material properties such as doping and trap concentrations, effective masses, and emission and capture rates.)

P570.2 Obtain the exact analytical formulae for the depletion-layer capacitance of a tunnel p/n junction, whose two quasi-neutral layers are both degenerate.

P581.1 Locate the limiting current points of the experimental s/m junction diodes given in Fig.563.1 based on the theoretical expression given in section 571. Take the work function values of the metals from Table 413.1, determine N_{DD} of each of the s/m diode in this figure.

P581.2 The accurate metal work functions are recently determined using the inverse problem of P571.1, that is, if N_{DD} is known precisely, the metal work function can be calculated from the experimental forward I-V curve of the m/s diode, provided that interface trap is negligible. Design an m/Si rectifying diode to measure the work function of Au in Si. (Hints: Determine the range of dopant concentration and dopant type to give a rectifying barrier, and determine the current density with a guessed work function for Au. Electron affinity of Si is given.)

P582.1 What is the limiting current density that can pass through the p+/n junction solved in problems P522.1, P523.1, P531.1 and P538.1 before ohmic drop becomes important? Assume a certain thickness for the Si wafer or the n-region and assume also perfect or zero-resistant ohmic contacts to both the p+ and n surfaces.

P583.1 Without using tunneling, design the surface concentration of the p+ and n+ surfaces of an m_1/p+/n/n+/m_2 diode with $m_1 = m_2 =$ aluminum so that the voltage drop across the two ohmic contacts are minimized. What metals or hypothetical metals (or metal work functions) are necessary if rectifying barrier persists?

P580.2 Use tunneling to improve the aluminum ohmic contacts in problem P580.1. **(a)** Obtain the tunnel current density formulae of the Al/Si triangular barrier from that of the SiO_2/Si triangular barrier given by (385.1A) using the theoretical formulae given by (154.1A). **(b)** Using tunneling, find the surface concentrations for the $m_1 =$ Al/p+ and n+/$m_2 =$ Al junctions in order to limit the ohmic drop to 1 mV across each of the ohmic contacts at the limiting (or flat-band) current density of P572.1. Disregard ohmic drop through the n-base layer.

P590.1 A perfectly crystalline (no defects) thin Si ($a_{Si} = 5.43072$A) epitaxial layer is commensurately grown on a Ge ($a_{Ge} = 5.65754$A) substrate to make a Si/Ge heterojunction. Thus, the Si layer is biaxially strained and tetragonally deformed (lattice constant: $a = b = a_{Ge} > a_{Si} > c$). Sketch the shift of the Si energy band and phonon spectra in the k_z (longitudinal or normal c-axis of the film) direction and in the film plane (k_x and k_y). Use the tight-binding or the plane-wave model or both and perturbation theory to give an estimate of the magnitude of the shift. Use Fig.183.1(b) for the bulk electron spectra and Fig.313.3(a) for the bulk phonon spectra in Si.

P590.2 Repeat problem P590.1 for a thin Ge film grown on a Si substrate.

P590.3 Using the Shockley diode formulae for a p/n homojunction, write down without algebra the diode current equation for heterojunction p/n diode with a valence discontinuity of ΔE_V and a conduction band discontinuity of ΔE_C.

P590.4 If mobilities are identical, what ΔE_G, ΔE_V and ΔE_C, are necessary to offset a ratio of N_{DD}/N_{AA} such that the minority carrier current on the two sides (two quasi-neutral layers) are equal? Assume both layers are thick compared with respective minority carrier diffusion length.

Chapter 6
METAL-OXIDE-SEMICONDUCTOR AND OTHER FIELD-EFFECT TRANSISTORS

The most abundant object made by man
on this planet earth!

600 INTRODUCTION 524
610 PHYSICAL STRUCTURE OF THE INVERSION-CHANNEL MOSFET 526
- Coordinate System, 526
- Body of the Semiconductor, 527
- Source and Drain, 527
- Gate Oxide, Gate Contact, Gate Width, 527
- Channel Type, Channel Length, Channel Thickness, 528
- Field Oxide, Pad Oxide, 528

620 QUALITATIVE DESCRIPTION OF DC MOSFET CHARACTERISTICS 529
621 Physics of MOST Channel Current, 529
622 Output and Transfer D.C. Characteristics, 529
623 Four Basic MOST Current-Voltage Characteristics, 530
630 TYPICAL FABRICATION STEPS OF AN N-CHANNEL MOSFET 533
- Description of the Fabrication Steps, 535

640 D.C. CHARACTERISTICS OF MOSFET (Elementary Analysis) 538
641 Conductivity Modulation Model of the MOSFET (D.C. Drift Current and Charge Control Analysis), 541
- Current-Charge Equation from Longitudinal Electric Field, 542
- Voltage-Current Equation from Transverse Electric Field, 545

642 Oxide and Interface Trapped Charges, (Instability, Aging, Failure), 547
643 The MOSFET Equations and D.C. Characteristics, 550
644 Numerical Examples of MOSFET D.C. Characteristics, 558
650 SMALL-SIGNAL EQUIVALENT CIRCUIT MODEL OF MOSFET 559
651 Charge-Control Analysis of Capacitance Elements, 560
- Asymptotic Results at Low Drain Voltage and Current Saturation, 565

652 High-Frequency Response of MOS Transistor, 568
- Transconductance Cutoff Frequency, 568
- Gain-Bandwidth Product, 570

653 Numerical Example of Small-Signal Characteristics, 571
654 Distributed Low-Frequency Small-Signal Model, 572

660 SWITCHING PROPERTIES OF MOSFET 576
- 661 Intrinsic Delay, 577
- 662 Power-Delay Product (Figure of Merit), 581
- 663 Charging and Discharging a Capacitor - Extrinsic Delays, 583
  - Fundamental MOSFET Switching Equation, 586
  - Charging a Capacitor, 588
  - Discharging a Capacitor, 590
  - Charging-Discharging Comparison, 592
  - Charging-Discharging Cycle Time, 592
  - Charge Transferring Between Two Capacitors, 592

670 CIRCUIT APPLICATIONS OF MOSFET 596
  - MOST Circuit Symbols - Evolution Chronology, 596
  - Current-Voltage Equations for MOS Circuit Analysis, 601
- 671 Dynamic Random Access Memory Cell, the DRAM, 603
  - Definition of Memory Terms, 603
  - Brief Manufacturing History of DRAM Chip, 605
  - Equivalent Circuit Model of a DRAM Cell, 608
  - Cell Array Architecture of a DRAM Chip, 608
  - Basic Operation Principle of DRAM Cell, 610
    - Write 1 Operation - Bit Bunching, Line Driver, Precharge, 611
    - Read 1 Operation - Destructive Readout, Self-Restore/Force-Read, Dummy Cell, 611
    - Refresh Cycle and Origin of 'Dynamic', 612
- 672 The MOS Inverter Circuits, 614
  - Encyclopedia of Twenty MOS Inverter Circuits, 614
  - Analysis of the Three NMOS Inverter Circuits, 620
    - The RENMOS Inverter, 621
    - The DENMOS Inverter, 622
    - The EENMOS Inverter, 623
    - Power Dissipation of the Three NMOS Inverters, 623
  - The CMOS Inverter Circuit, 624
    - Why CMOS?, 624
    - Evolution History of CMOS Structures, 625
    - D.C. Analysis of the CMOS Inverter Circuit, 629
- 673 Static Random Access Memory Cell (SRAM) 632
- 674 Nonvolatile Random Access MOS Memories (ROM,PROM's), 636
  - Read Only Memory (ROM), 636
  - Programmable Read Only Memory (PROM), 637
  - Erasable Programmable Read Only Memories (EPROMs), 637
    - UV-EPROM (1971-1991!), 639
    - Flash-EPROM, 640
    - F-EPROMs (FMOS and FRAM), 641

680 BEYOND THE CONDUCTIVITY MODULATION MODEL 644
- 681 Effect of bulk charge (Body Effects: Dopant and Substrate Bias), 645
  - Origin of the Bulk Charge, 645
  - Analytical Approximation of Bulk Charge (Depletion Model), 647
  - Body Effects on Threshold Voltage, 649
  - Body Effects on I-V Shape, 650
  - Five Illustrations and a Numerical Example of the Body Effect, 651
- 682 Subthreshold Characteristics (Diffusion Current), 653
  - Definition of the Subthreshold Range, 656
  - Intrinsic Surface - Onset of Drift Current, 656
  - Analysis of the Subthreshold Current, 657
  - Subthreshold Drain Current Equation, 660
  - Dependence of Subthreshold Current on Drain Voltage, 660
  - Dependence of Subthreshold Current on Gate Voltage, 660
  - Temperature Dependence of the Subthreshold Current, 661
- 683 Effects of Oxide and Interface Traps, 662
  - Oxide Traps, 662
  - Interface Traps, 665
- 684 High Electric Field and High Voltage Effects, 667
  - I-V with Electric Field Dependent Mobility, 668
    - Longitudinal Field Dependent Mobility Effects, 669
    - Transverse Field Dependent Mobility Effects, 670
  - Electric Field and Voltage Dependent Generation-Recombination-Trapping, 674
    - Five Trap Charging-Discharging Processes via Capture and Emission, 675
    - Ten Trap Charging-Discharging Processes via Elastic Tunneling, 676
    - Eight Trap Generation-Annihilation Processes via Hydrogenation, 676
    - Hydrogenation of Boron and Group-III Acceptors in Si, 677
- 685 Short Channel and Narrow Gate Effects, 679
  - Three Short Channel Effects and Lowly Doped Drain, 679
  - Three Narrow Gate Effects, 683

690 OTHER FIELD-EFFECT TRANSISTORS - EVOLUTION HISTORY 687
- MESFETs, 687
- MOSFETs, 687
- The First JGFET (Surface-Ion Induced Inversion Channel), 689
- JGFET, 690
- Two Current Saturation Mechanisms (Channel Pinch-off and Carrier Depletion), 690
- High-Mobility Confined-Channel Heterojunction FETs, 691

699 BIBLIOGRAPHY AND PROBLEMS 694

600 INTRODUCTION

The metal-oxide-semiconductor field-effect transistor, **MOSFET**, is a four-terminal semiconductor device. The main feature of its electrical characteristics comes from modulation of the conductivity of a semiconducting resistor by a voltage or electric field applied **perpendicular** to the length of the resistor. This is known as the **conductivity modulation** or **field-effect principle**.

The fundamental properties of semiconductor material described in chapters 1 and 2 and the mechanism of the drift current described in sections 310-314 are sufficient to understand the physics and to derive the equations of the main range of the d.c. current-voltage characteristics of the MOSFET. In the main range, diffusion and generation-recombination-trapping currents are not important compared with the drift current. The mathematical analysis is further simplified because both the low-frequency a.c. equivalent circuit and the switching speed of the intrinsic MOSFET can be derived using the charge-control method and studied using elementary circuit analysis concepts. The theory predicts the experimental data quite well except when the MOSFET operates at very low currents or very high voltages, or when it is very small.

For operation at very low currents, diffusion current in the MOSFET channel dominates over drift current. Thus, diffusion must be taken into account. The current-voltage characteristics at low currents is similar to that due to injection and diffusion of minority carriers in a p/n junction.

At very high voltages, generation of electron-hole pairs by the interband impact mechanism described in section 3613 and applied to a reverse biased p/n junction in section 536 must be considered. This mechanism determines the maximum voltage of operation of a MOSFET. It also affects the stability of the electrical characteristics and the operating life because the energetic or hot semiconductor electrons that are generated in the high electric field from the high voltage can climb over a thick or tunnel through a thin gate-oxide/silicon triangular potential barrier. Once the semiconductor electrons are in the gate oxide layer, they can charge up the oxide if they are captured by the oxide traps. The trapped oxide charges cause instability because it shifts the current-voltage characteristics. Furthermore, the volume density of the trapped oxide charge may become so high at the trap sites as to cause a high local current density which would heat up and burn holes in the oxide film. These effects will be briefly discussed in this chapter.

For very small MOSFET's, two-dimensional and three-dimensional effects become important even if the electric field is not high. In high frequency and high speed applications, parasitic capacitances also become important. We will not be able to treat all the two-dimensional effects adequately in this introduction text. However, the parasitic capacitances and some two- and three-dimensional effects can be analyzed approximately by partitioning the transistor structure and analyzing each partition one-dimensionally.

There are many acronyms for the MOSFET, such as, the metal-oxide-semiconductor transistor, **MOST**, first used by Wanlass and Sah in the initial draft of the 1962 conference paper that disclosed the complementary MOS transistor circuit (now known as **CMOS**), the metal-insulator-semiconductor field-effect transistor, MISFET (a misfit), the metal-oxide-semiconductor field-effect transistor, **MOSFET**, and others. The acronym **MOSFET** appeared to have been used first by engineers at the Fairchild Semiconductor Corporation in the early 1960's. It was later accepted by IBM researchers, and is almost universally adopted now. But for simplicity and consistency with the terminology and acronym of the other major semiconductor device - the bipolar junction transistor (BJT) - we frequently use the term **MOS Transistor** or the acronym, **MOST**, in place of MOSFET. I wrote a comprehensive invited review on the evolution history and future prospects of the MOS transistor in the October 1988 issue of the Proceedings of the IEEE which may be a source for further reference [600.1].

[600.1] Chih-Tang Sah, "**Evolution of the MOS Transistor - From Conception to VLSI**," Proceedings of the IEEE, 76(10), pp.1280-1326, Oct. 1988.

The dominant d.c. current in the MOSFET is carried by only one type of charge carriers, either electrons (**n-MOST**) or holes (**p-MOST**). Thus, it is historically known as an **unipolar transistor**. Sometimes it is also known as a majority carrier device but this is incorrect since there are two types of MOSFET, one operates with majority carriers and the other, minority carriers. The majority-carrier MOSFET operates with majority carriers (say, electrons) from a chemical impurity (phosphorus, arsenic, or antimony) that is intentionally incorporated into a thin surface layer to give the n-type conduction channel on the p-type Si substrate. This is known as the **doped-channel MOSFET**. The minority-carrier MOSFET operates with minorities carriers (say, electrons on a p-type substrate) '**injected**' from a n-type source region into a surface conduction channel and collected by a n-type drain region. The surface channel is **inverted** to n-type if its electron concentration becomes higher than its hole concentration. This is known as the **inversion-channel MOSFET**. The n-type surface inversion channel on the p-type Si substrate can be induced by either an electric field or voltage applied perpendicular to the surface or by imbedding a layer of positive charge in the surface oxide film. In contrast, the bipolar transistor current depends on the presence of both electrons and holes or both minority and majority carriers so the name **bipolar junction transistor** (**BJT**). BJT is sometimes known as a minority-carrier device because its principal output current component is carried by injected minority carriers which pass through a thin base layer and are collected by a reverse-biased p/n junction in the output.

The historical names and the traditional explanations given in many textbooks and journal articles are often inaccurate, sometimes incorrect, and infrequently bring out these key fundamental differences and similarities between the unipolar and bipolar transistors. As we continue in our exposition and study of the electrical characteristics, we will obtain an in-depth understanding of the physical basis of the differences and similarities.

610 PHYSICAL STRUCTURE OF THE INVERSION-CHANNEL MOSFET

We shall first give a description of the structure of a MOSFET in order to delineate the physical and geometrical regions of the device and to define the electrodes and terminals. The n-type inversion-channel Si MOSFET (nMOST n-MOST) shown in Fig.610.1 will be used as the model example.

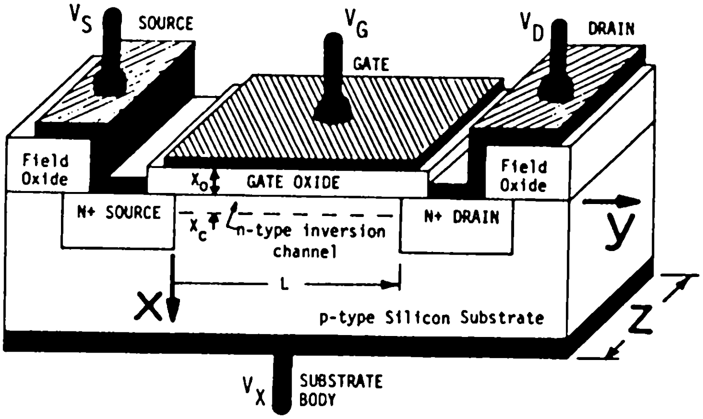


Fig.610.1 The three-dimensional view of inversion n-channel Si MOST model example for defining the nomenclatures, physical dimensions and electrical parameters described in the text.

Coordinate System

Although we frequently use the circular or ring-dot geometry to describe a MOST (center dot is the drain and rings are source and gate), the more common geometry used in VLSI chips is rectangular as shown in Fig.610.1. The x-axis denotes the **depth direction** into the Si substrate. The y-axis denotes the **channel length direction** along which the current flows, known also as the **longitudinal direction** since it is longitudinal to current flow. Thus, the x-direction in the oxide film is the direction of the **transverse** electric field that controls the current, i.e., the oxide field (x-direction) is transverse to the current (y-direction). The z-axis labels the **width direction** both for the **gate width** and the **channel width**.

Body of the Semiconductor

The semiconductor **body** of the n-inversion channel MOST shown in Fig.610.1 is the p-type Si substrate. It is also known as the semiconductor **bulk**, **substrate** or **base**. The darkened layer at the lower surface in the figure is the conductor electrode which makes an ohmic contact to the Si body and it is connected to a terminal to the external circuits. It is known as the **body electrode**. The d.c. voltage applied to the lead or wire connected to the body electrode is denoted by V_X (from x-axis) since subscript B (for body) is used for the base of BJT or the breakdown voltages and S (substrate), for the source voltage.

Source and Drain

The two N+ regions in Fig.610.1 form the two n^+/p junctions. The '+' in N+ and n^+ means that the n-type region is heavily doped and has very high conductivity or low resistivity. The N+ regions were coined as the **source** and the **drain** by Shockley when he invented the junction-gate field-effect transistor in 1952. Shockley's basis for these choices was that the N+ source region is the source of charge carriers that flow in the channel (electrons in the nMOST) while the n-type drain region is the sink or the drain of the charge carriers. Shockley discarded the term 'sink' in order to avoid subscript conflict with source and substrate.

For a symmetrical MOSFET, the two N+ regions and their adjacent gate oxide are identical so that the source and drain are interchangeable. Frequently, the source and drain junction areas are different and the adjacent gate oxides contain different densities of oxide and interface traps created by the different high electric fields near the source and drain junctions. In special circuits, such as the MOSFET switch in the DRAM (dynamic random access memory) cell, the source or drain region may even be missing. For these reasons and for expediency in circuit analysis, the source and drain regions are labeled explicitly as S and D.

The surface of the two N+ regions is contacted by a conductor which is shaded in solid dark in Fig.610.1. These are known as the source and drain **electrodes** or **contact pads** and the leads or wires attached to these electrodes or pads are known as the source and drain **terminals**.

The d.c. voltages applied to the source and drain are labeled V_S and V_D respectively. The d.c. current flowing into the drain electrode from the drain terminal wire is labeled I_D.

Gate Oxide, Gate Contact, Gate Width

The third region of the MOSFET is the surface oxide layer between the source and drain. This is a thin layer of pure, defect-free, and 50-2000A thick thermally grown oxide. It serves as the dielectric layer so that the gate can sustain as high as 1×10^6 to 5×10^6 V/cm (1 to 5 MV/cm) transverse electric field in order to strongly modulate the conductance of the channel. This is known as the **gate oxide**.

Above the gate oxide is a thin gate electrode layer made of a conductor which can be aluminum, a highly doped silicon, a refractory metal (tungsten), a silicide (TiSi, MoSi, TaSi or WSi) or a sandwich of these layers. This **gate electrode** is often called the **gate metal** or **gate conductor**. The voltage applied to the gate electrode is labeled V_G. The **geometrical width** of the gate conductor electrode, Z, is called the **gate width**. In current practice, the symbol W (for width) is frequently used instead of Z. Z or W is slightly smaller than the gate width defined on the lithographic mask due to undercutting during chemical etching to define the gate electrode during lithography. The gate width or **geometrical gate width** is not to be confused with the conduction channel width or **electrical channel width** since they are not equal. The conduction channel may be slightly wider due to fringe electric fields at the two width (Z-direction) edges of the gate electrode.

Channel Type, Channel Length and Channel Thickness

Below the gate oxide is a thin n-type inversion layer on the surface of the p-type Si substrate. It is induced by the oxide electric field from the applied gate voltage, V_G. This is known as the **inversion channel** or more specifically, the **n-type inversion channel** or just **n-inversion channel**. It is the conduction channel that allows the electrons to flow from the source to the drain. The **geometrical channel length** is labeled by L. The electrical channel length and width, L_E and W_E, are the parameters that enter the MOST equations, not the geometrical channel length and geometrical gate width. Due to the invasive electric field and the junction space-charge layer from the reverse biased P/N+ drain junction, the electrical channel is shorter than the geometrical channel. The **electrical channel thickness** is labeled by X_c or x_c. In a built-in or doped n-channel MOST, this is equal to the geometrical thickness of the doped channel or the depth of the n/p+ junction of the doped layer minus the thickness of the junction space-charge layer on the doped n-channel side (instead of the p-substrate side).

Field Oxide, Pad Oxide

The oxide layers under the the source and drain contact bonding pads, about 5000A, are much thicker than the gate oxide of 50A-2000A. The thicker oxide reduces the loading capacitance and increases the dielectric breakdown voltage to allow the drain to be biased to higher voltages. The oxide at the two width edges (Z-direction) of the gate is also thicker (not shown in Fig.610.1). These thick oxides, surrounding the thin gate oxide and covering the drain and source junctions, are known as the **field oxides** and those under the wire bonding metal (Al) pad are also known as **pad oxides**.

620 QUALITATIVE DESCRIPTION OF MOSFET DC CHARACTERISTICS

We shall first give a qualitative description of the physics and origin of channel current and its modulation. From this, a catalog of MOSFET structures and operation conditions is obtained. These are given in the following two sections.

621 Physics of MOST Channel Current

The operation of a field-effect transistor is based on the modulation of the conductance of a resistive channel by a transverse electric field. This can be illustrated using the inversion n-channel MOSFET shown in Fig.610.1. When there is no voltage applied between the gate and the substrate (body or bulk) terminals, there is no surface conduction path connecting the n-type source and drain regions. The only current path between the source and drain is through the two n+/p junctions and the p-type bulk. But the two n+/p rectifying junctions are connected back-to-back so that one of the n+/p junctions will be reverse biased and passes very little current when a d.c. voltage is applied between the drain and source. This is the **off region** or **cutoff range** of the MOSFET current-voltage characteristics. It is also known as the **cutoff mode** during circuit operation of the MOST.

When a positive voltage is applied between the gate and the substrate, electrons are attracted to the oxide/silicon interface under the gate electrode to form an n-type surface inversion layer, labeled by 'n-type inversion channel' in Fig.610.1. This n-type layer forms a conduction channel which connects the n-type source and drain regions. When a d.c. voltage, V_D, is applied between the source and drain terminals (for n-channel, $V_D > 0$), electrons will flow from the source to the positively biased drain. A corresponding d.c. current will flow into the drain terminal from the external circuit and out of the source terminal into the external circuit.

622 Output and Transfer D.C. Characteristics

A family of d.c. drain current versus d.c. drain voltage (I_D-V_D) characteristics of a typical n-inversion channel Si MOSFET is shown in Fig.622.1(a). The gate voltage is the constant parameter. This is known as the I_D-V_D plot and the **output characteristics**.

Note that the drain current **saturates** to a constant value which is independent of the drain voltage when the drain voltage is equal to or greater than the gate voltage. The saturated drain current values are shown by thick dark lines. This region of the I_D-V_D characteristics is known as the **saturation region**. It is sometimes also known as the **pentode region** after the pentode vacuum tubes which have a similar saturated current-voltage characteristics.

The other region, above the dashed parabola and under the thin light concave curves shown in Fig.622.1(a), is known as the **non-saturation region**. It is also known as the **triode region** after the triode vacuum tubes which have a similar non-saturated but convex current-voltage characteristics.

The d.c. characteristics can be displayed in a second way, plotting the drain current as a function of gate voltage with the drain voltage as the parameter. This is shown in Fig.622.1(b) and is known as the **transfer characteristics**. The lighter lines are straight lines corresponding to the non-saturation region of the output characteristics shown in figure (a). The thick, dark and curved portion of the transfer characteristics corresponds to the saturation region of the output characteristics shown in figure (a). The dark curve is actually a true parabola in the idealized intrinsic MOST model we will analyze in sections 64n. The transfer characteristic is particularly useful to help categorize four types of MOSTs described in the next section.

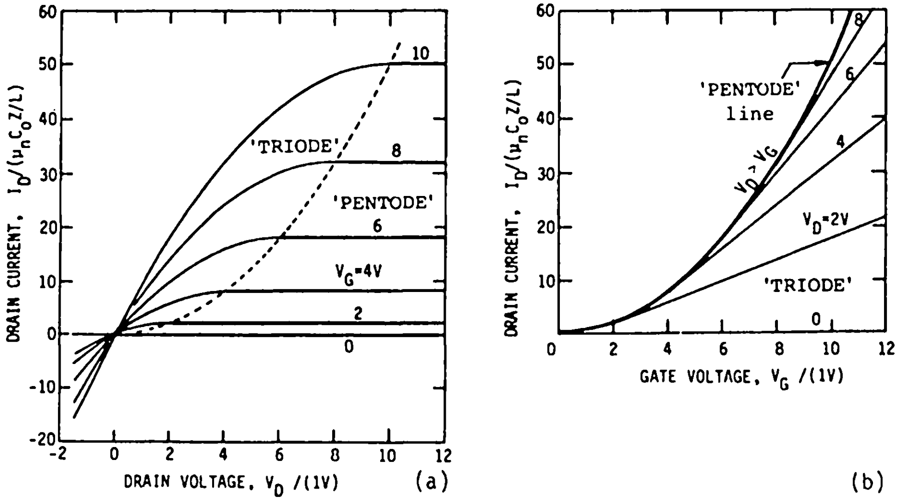


Fig.622.1 Typical d.c. characteristics of an inversion n-channel MOSFET. (a) The output characteristics. (b) The transfer characteristics. Note normalized current unit is volt^2.

623 Four Basic MOST Current-Voltage Characteristics

For both fundamental understanding and expedient circuit analysis of the MOST, the two channel conductivity types (n-type and p-type) have been divided into two **physical origins**, giving four basic types of MOSTs and four families of current-voltage characteristics. MOSTs have also been described by two **operation modes**. The physical origins and operation modes are inter-related. A qualitative description of the origins and modes and a definition of terms are given in this section to facilitate future discussions.

The two physical origins are the **induced channel** and the **doped channel** distinguished by the presence and absence of impurity doping to create the channel. Thus, there are four types of MOST, the induced or doped n- and p-channels. The two operation modes refer to the **absence (enhancement mode)** and **presence (depletion mode)** of a conduction channel **at zero applied gate voltage**. These terms were introduced by S. R. Hofstein and F. P. Heiman [Proc. IEEE, **52**, 1190(1961)] who did not consider the channel induced or depleted by oxide-interface charges described below.

The device structures of the two physical-channel origins are shown in Figs.623.1(a)-(b) for the n-channel and (c)-(d) for the p-channel MOSTs. Figure 623.1(e) uses the transfer characteristics (I_D-V_G) to illustrate the operation of the induced-channel nMOST and pMOST in the enhancement mode, while Fig.623.1(f) illustrates the operation of the doped channel nMOST and pMOST in both the depletion and enhancement mode. In the following paragraphs, the n-channel MOST is described to illustrate these physical origins and operation modes.

The differences between the induced- and doped-channel nMOST, shown in Figs.623.1(a) and (b), lie in the way the surface conduction channel is created. Consider first the **induced** n-type **inversion channel** on the p-Si substrate. Its n-type surface conduction channel is induced by an electric field perpendicular to the Si surface covered by the gated SiO_2. This electric field attracts the electrons (minority carriers) to the Si-surface from the p-Si bulk and from the N+ source and drain regions to form the n-type surface channel. Simultaneously, it also repels the holes from the surface. This electric field can be created by **(1)** a voltage applied to the gate over the SiO_2, or **(2)** by a built-in potential drop across the gated SiO_2 and the semiconductor surface layer. The built-in potential drop can come from **(2a)** an electric dipole layer due to the work function difference between the gate conductor and the semiconductor or **(2b)** a layer of positive charge trapped (imbedded) in the oxide layer or at the oxide/semiconductor interface. Both of these, if sufficiently large in magnitude, would give an existing (i.e. **built-in**) **induced inversion channel** at zero bias voltage.

In contrast, the **doped** inversion **n-channel** on a p-Si substrate shown in Fig.623.1(b) has an existing (**built-in**) n-type semiconductor surface layer between the source and the drain at zero bias voltage from donor impurity doping. This n-type **built-in doped-channel** is produced by introducing a controlled amount of donor (P, As or Sb) impurity into the surface layer during Si epitaxial growth or via ion implantation. Impurity diffusion is not used since it cannot precisely control the very low donor impurity concentration required to pinch off the doped channel.

The electrical differences between the two channel origins and the two modes are most easily illustrated by the transfer characteristics shown in Figs.623.1(e) and (f). Consider the nMOST curve. In general, there is a gate voltage above which the surface channel is present and drain current begins to flow. Below this gate voltage, the channel is absent and the drain current ceases. This gate voltage is labeled V_T, V_{TH} or V_{GT}. It was coined the **threshold voltage** by this author in 1964. For the nMOST, it is labeled V_{TN} and for the pMOST, V_{TP}.

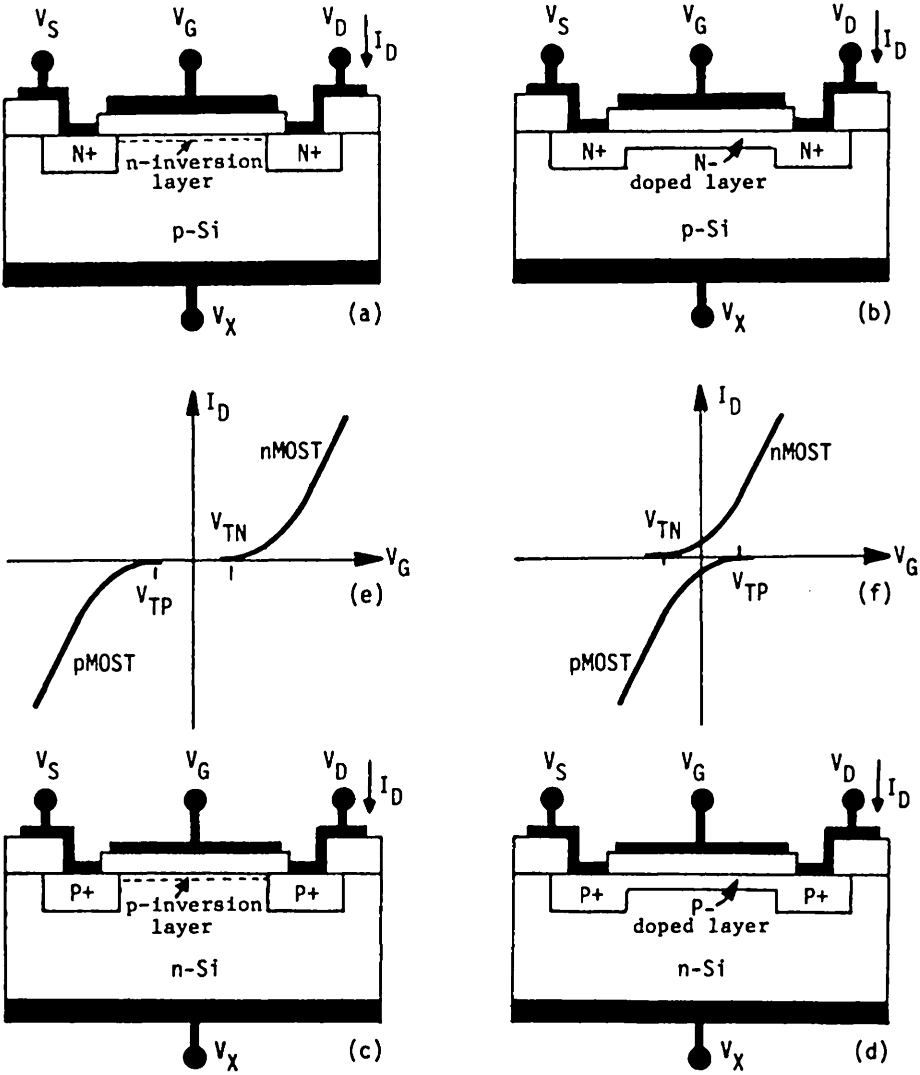

V_S
V_G
V_D
I_D
N+
n-inversion layer
N+
p-Si
V_X
(a)
V_S
V_G
V_D
I_D
N+
N-doped layer
N+
p-Si
V_X
(b)
I_D
nMOST
V_{TN}
V_G
V_{TP}
(e)
pMOST
I_D
nMOST
V_{TN}
V_G
V_{TP}
(f)
pMOST
V_S
V_G
V_D
I_D
P+
p-inversion layer
P+
n-Si
V_X
(c)
V_S
V_G
V_D
I_D
P+
P-doped layer
P+
n-Si
V_X
(d)

Fig.623.1 The two different structural or physical origins and two operation modes of MOSFETs. (a) The n-induced channel. (b) The n-doped channel. (c) The p-induced channel. (d) The p-doped channel. (e) The transfer characteristics of the enhancement-mode nMOST and pMOST. (f) The transfer characteristics of the depletion-mode nMOST and pMOST.

For the induced n-channel illustrated in Fig.623.1(e), V_{TN} is positive. This means that the n-channel does not exist at zero gate voltage and a positive gate voltage is necessary to create the n-channel and start a drain current. This is known as the enhancement-mode MOSFET, that is, an applied gate voltage is required to enhance the minority carrier density at the semiconductor surface in order to create an inversion channel and to start the drain current.

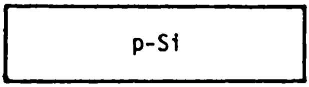
For the doped n-channel illustrated in Fig.623.1(f), there is already a built-in channel at zero gate voltage to enable a channel or drain current to flow between the drain and the source. Thus, a negative gate voltage, $V_{TN}<0$, is necessary to deplete the carriers in the channel in order to cut off conduction and drain current. This transistor is known as the depletion-mode MOSFET.

It is obvious that the use of the terms enhancement-mode and depletion-mode to classify the MOS transistors is ambiguous which was not recognized by Hofstein and Heiman in 1963 since they did not consider a **built-in inversion channel** induced or depleted by a layer of oxide and interface charges. For example, the depletion-mode nMOST with the doped channel shown in Fig.623.1(f) can also operate in the enhancement mode by merely applying a positive voltage to the gate. Similarly, the enhancement-mode nMOST in Fig.623.1(e) can operate in the depletion mode relative to some steady-state positive gate voltage, say $V_G=10V > V_{TN}=5V$, if a negative 2V gate voltage pulse is superimposed onto the 10V d.c. gate bias to reduce the n-channel conductivity. One can also have more complicated ambiguities. For example, an n-type inversion channel is induced by a layer of positive charge in the oxide. Then, there is a conduction channel and drain current at zero gate voltage. This would be a depletion-mode inversion-channel nMOST. For a further example, the doped-channel in an nMOST is pinched off by a layer of negative oxide charge. Then there is no conduction channel between drain and source at zero gate voltage. This would be an enhancement-mode doped-channel nMOST.

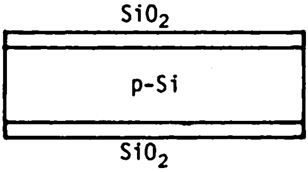
Thus, the appropriate **device classification** are the **two channel conductivity types** and the **two channel physical origins**. The depletion and enhancement terms should be strictly limited to describing the **mode of operation** in a circuit. They should not be used to classify the MOST devices. Unfortunately this has not been followed by device and circuit design engineers. Furthermore, the induced inversion-channel MOST is often imprecisely and incompletely termed the inversion channel MOST (a usage originated by Bell Lab researchers and inventors). Nevertheless, in later sections we will use the term inversion channel to mean an induced inversion channel MOST.

630 TYPICAL FABRICATION STEPS OF AN N-CHANNEL MOSFET

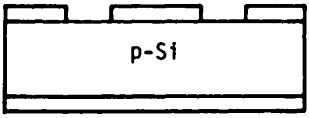
A typical sequence of steps to fabricate an n-channel MOSFET is shown in Fig.630.1 and described next, starting with a single crystal silicon slice.



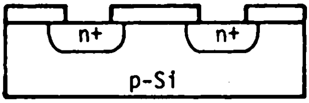
(1) Etch and clean both surfaces.



(2) Grow oxide layer on both surfaces.

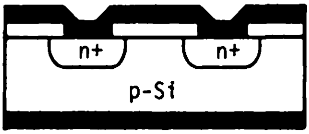


(3) Etch source and drain diffusion windows on top surface using photolithography.

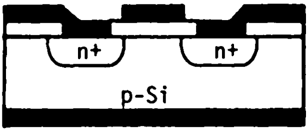


(4) Diffuse or ion implant donor impurity (P, As or Sb) into the oxide windows to form source and drain n+/p junctions.

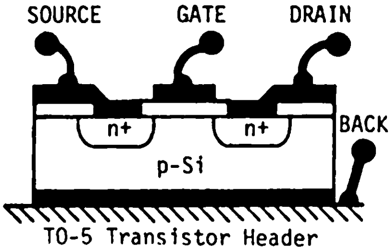
(5) Remove oxide layer on bottom surface.



(6) Blanket evaporation of pure aluminum layer on both top and bottom surfaces.



(7) Define contact area of gate, drain and source by etching aluminum using photolithography.



(8) Dice wafer.

(9) Solder die onto a 4-pin gold-plated transistor header (such as TO-5).

(10) Bond 1-mil gold wire between the pin-posts and the source, gate, drain metal contact areas and the header.

Fig.630.1 Ten typical steps required to make an n-channel silicon MOSFET.

Description of the Fabrication Steps

(1) Chemically etch and clean both surfaces of a p-type silicon slice to remove surface damage and contamination. The silicon slice is about 500 micron or 20 mil thick for silicon disks of four to eight inches in diameter used in today's production. The terms wafer, disk, and slice are used interchangeably for the starting Si slice. Si wafer and Si disk are used since the starting Si piece is often circular while the term 'slice' suggests a square or rectangular shape similar to that of a slice of bread when the starting Si wafer is pre-shaped into a rectangular or square piece. In early days (late 1950s and early 1960s), the Si wafer was about 250 micron or 10 mil thick and one to two inches in diameter. It has been increased to 15cm (6 inch) or 20cm (8 inch). The bulk resistivity of the Si slice may be 1 to 10 ohm-cms. For higher speed devices, the Si bulk resistivity may be lower or the Si slice may have a thin (5-10 microns) p-type epitaxy layer of 50 ohm-cm grown on a 0.01 ohm-cm p-type Si substrate. Thicker and higher resistivity epitaxy layers and higher resistivity substrates are required for high voltage and high power MOS transistors and MOS integrated circuits.

(2) Grow an oxide film of 2000A for diffusion mask on both surfaces in dry and pure oxygen at about 1000°C in a very clean furnace tube made of highly pure fused silica (SiO_2 glass). For future high-speed submicron VLSI MOSTs, the gate oxide must be decreased to less than 100A thick. An etch and regrowth step will be used to grow such a thin oxide. The procedure consists of etching off the 2000A oxide, then regrowing the oxide at a lower temperature (800°C) after the n+ source and n+ drain are formed in the following paragraphs.

(3) Etch the source and drain diffusion windows in the oxide layer of the chemically polished and oxidized upper or top surface of the Si wafer. The size and shape of the window are defined by photolithography. Future high-speed submicron MOSTs and BJTs may use x-ray or electron-beam lithography in place of photolithography since the transistor dimensions will be smaller than the wavelength of visible light (purple is about 0.35 μm or 3500A). X-ray and E-beam lithography would produce radiation damage, which are weakened (strained), or ruptured (dangling) Si-Si and Si-O bonds in the oxide and at the oxide/silicon interface. The dangling bonds are electron and hole traps. Thus, in order to maintain performance and give an operating life greater than 10 years, the weakened and dangling bonds must be healed at a high temperature ($<\sim$900C) in a dry hydrogen-free ambient to avoid formation of hydrogen bonds which are also weak.

(4) Diffuse phosphorus (or As or Sb) impurity through the windows of the oxide film into the p-type silicon to form the n+ source and drain regions and n+/p isolation junctions. The oxide will mask the p-Si substrate so that the phosphorus (or arsenic) donor impurity can only diffuse into the p-Si through the windows in the oxide where the oxide was removed in (3). The diffusion is carried out between 800C and 1000C, in an ambient of dry nitrogen with a low partial

pressure of a phosphorus-containing gas such as $POCl_3$ (phosphorus oxychloride), P_2O_5 (phosphorus pentoxide), PBr_3 (phosphorus tribromide), PCl_5 (phosphorus pentachloride), PH_3 (phosphine), or AsH_3 (arsine) and other P, As, Sb compounds. P_2O_5 was used in the first impurity diffusion experiments in Si undertaken by Fuller and Ditzenberger in 1954 at Bell Telephone Laboratories [600.1] and in the first production of diffused silicon diodes and transistors in 1959 at Fairchild . Solid silicon-nitride disk impregnated with phosphorus is the choice in today's production lines because it avoids the messy handling requirements of the toxic, reactive, incinerative, and explosive gas and liquid sources. Areal uniformity of the impurity concentration is also better from the solid silicon nitride disk source. However, the phosphorus doped nitride disk must be larger than the 6-8 inch silicon wafer and at least one nitride disk is required for two Si wafers. The wafer boat needed to carry 100 8-inch Si disks and more than 50 nitride disks is large and heavy. Thus, ultraclean mechanical handling apparatus and techniques (robotic) are crucial in manufacturing to achieve high yield and low cost. Alternatively, phosphorus (or As, Sb) ion implantation can be used to give the n+ source and drain which would require a post-implant heating between 900C-1000C to remove the damage and to activate and diffuse-in the phosphorus donors.

(5) The top surface of the Si disk is then protected by a coating of clean and chemically inert photoresist film and the oxide layer on the back surface is etched off. There is an additional thin oxide layer on the top surface which is formed during phosphorus diffusion because of some oxygen in the furnace tube from the phosphorus source. This thin oxide has a high phosphorus concentration and is known as the **phospho-silicate glass** (PSG). It is a high-concentration phosphorus source and must also be etched off. If left on the Si surface, it may produce undesirable n-type layer on the exposed surface regions of the p-type Si via vapor transport. Furthermore, the high phosphorus concentration from PSG may also cause the phosphorus atoms to cluster in the n+diffused source and drain regions which could electrically short circuit the source and drain n+/p junctions.

(6) A thin layer of ultra pure aluminum (99.999% purity or better) of about 5000A is then evaporated onto both the top and the bottom surfaces in a high vacuum station at a pressure lower than 10^{-6} torr. The aluminum evaporation boat, usually made of tungsten, is precleaned by electrical heating to white heat (or a temperature much higher than the Al melting and evaporation point, 585°C) in the vacuum before the Al is loaded. This would drive off all the impurities in the boat, in particular, sodium, which is a very mobile positive ion in the gate oxide and is the main source of electrical instability of the MOSFET. A shutter is also used to allow the impurities in the Al to evaporate onto the shutter so that only truly pure Al is evaporated onto the oxidized Si surface after all the impurities in the Al have evaporated onto the shutter. Shuttering was the most crucial manufacturing trick invented by a Fairchild team (Deal, Grove, Snow, Sah) under this author's direction in 1962 to prevent sodium contamination and give stable MOSFET. It is still used today since it is impossible to keep the aluminum free from the minute trace of salt

(NaCl) in the ambient air while transporting the aluminum or opening the vacuum evaporation system to the ambient during loading and removal of the Si wafers.

(7) Define the contact electrode patterns of the source, drain and gate via wet or dry (gaseous) etching through windows defined by photolithography. The aluminum film on the Si wafer's back surface is protected during this etching step.

(8) There are now many MOSTs on a 4-8 inch silicon wafer. Only one transistor is shown in the diagrams just discussed. The many transistors on the wafer are separated by dicing the Si wafer into small squares or dice of about 50x50 to 100x100 mil^2 with a diamond impregnated high-speed cutting wheel. Each die contains one transistor. The die is popularly known as the **microchip** or **chip**.

(9) A die is then soldered onto a 4-pin gold-plated TO-5 transistor header (diameter ≈ 3/8") shown in the three-dimensional view of a finished device in Fig.630.2. This operation is known as **die attach** or **die bonding**. Gold plating of the header improves the alloying of the die to the header. It also helps the following wire bonding step to give strong mechanical bond.

(10) Gold wires of 1 mil diameter are attached to the aluminum contact electrodes of the gate, source and drain by ultrasonic or thermo-compression bonding. The other end of these wires is bonded to the three gold-plated metal posts of the 4-pin transistor header. The fourth post is wire bonded to the metal platform of the transistor header which is in electrical contact with the substrate when the die is soldered onto the header. This completes a typical sequence of steps to make a single canned MOST.

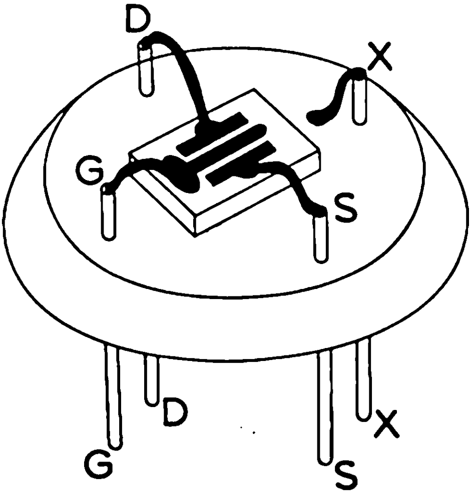


Fig.630.2 An Si nMOST chip mounted onto a 4-pin TO-5 header.

640 D.C. CHARACTERISTICS OF MOSFET (Elementary Analysis)

Three linear geometries for n-channel MOST are shown in Figs.640.1(a) to (c). The critical device dimensions are labeled in these figures which are the oxide thickness, x_o, the channel thickness, x_c, the channel length, L, and the gate width, Z, or channel width, W, used interchangeably later. These dimensions determine the current-voltage-power characteristics, the maximum switching speed and the highest frequency of oscillation. The switching speed and the small-signal high-frequency response are degraded by the drain n+/p junction capacitance, and by external loading from stray or parasitic capacitances and inductances of (i) the electrodes and (ii) the connections employed to interconnect the MOSTs with each other and with other devices and circuits on and off the Si chip.

Figure 640.1(a) shows a small-geometry, high-speed and low-power MOST. It illustrates the geometrical proportion of recent and future submicron MOSTs used in high-density chips of very-large-scale, ultra-large-scale, and extremely-large-scale integrated circuits (VLSI, ULSI, and ELSI circuits). In this MOST, $Z \sim L > x_o >> x_c$. The inequalities have the following meaning.

$Z \sim L$	means that the gate width, Z, is about the same as the channel length, L.
$L > x_o$	means that the channel length, L, is greater than the oxide thickness, x_o, by a factor of about two or three.
$x_o >> x_c$	means the oxide thickness, x_o, is much greater than the channel thickness x_c or more than 10 times greater, $x_o > 10x_c$.

In an analysis where an approximation can be made because of the presence of a large difference between the values of two parameters, usually the range of the two parameters and the error introduced by the approximation are stated. These stated conditions quantify the qualitative discussion of the results.

The selection of the relative size of these dimensions for a small-geometry, high-speed and low-power MOST comes from the following considerations. The short channel length is necessary to give short transit time of electrons drifting through the channel from the source to the drain. The transit time is about the shortest switching time the MOST can attain. The reciprocal transit time is about the highest frequency the MOST can still oscillate or give voltage amplification. The narrow gate (sometime incorrectly called narrow channel since it is the gate width that defines the modulated channel width) or small Z is used to reduce the cross sectional area of the channel so that the channel or drain current is low in each MOSFET. Then, the power density (power/area) or total d.c. power used by a million MOSTs on a VLSI chip can be dissipated by the chip carrier and package to maintain an acceptable chip and transistor temperature. The thin oxide x_o and thin channel x_c are needed to give high transconductance and gain in switching and amplifying circuits. The choice of $L > x_0 >> x_c$ is necessary to retain the current saturation characteristics shown in Fig.622.1(a) which is highly desirable in circuit applications. This is known as the **scaling law**. It was first proposed and demonstrated by Dennard of IBM in 1973 [600.1].

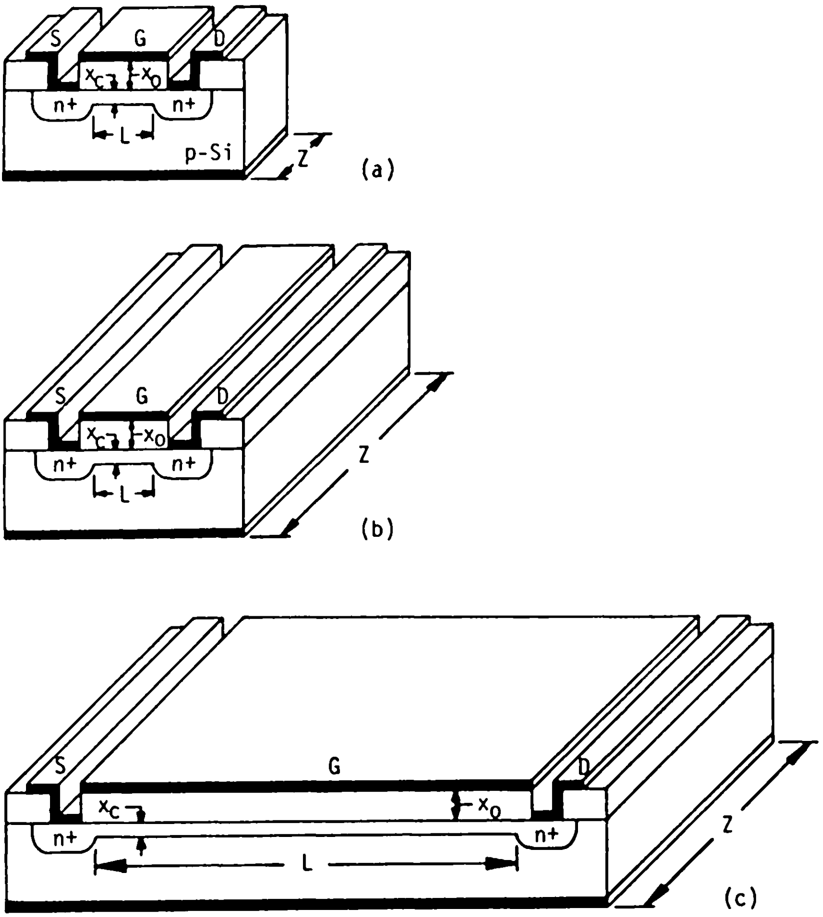


Fig.640.1 Geometry of three n-channel MOSFET's. (a) Narrow-gate, short-channel, high-speed, and low-power. (b) Wide-gate, short-channel, high-frequency, high-speed, medium-current and low-voltage. (c) Wide-gate, long-channel, medium-speed, medium-current, high-voltage, and medium-power.

Figure 640.1(b) shows a high-frequency or high-speed and medium-power or high-current and low-voltage MOST. The short channel length, L, is again required to give small electron transit time through the channel for high-frequency and high switching-speed operations. The channel or gate is very wide compared with the channel length, channel thickness, and oxide thickness. The wide channel gives a large cross sectional area to provide the high drain current to drive a large capacitance load.

For medium voltage operation (V_D > about 20V), the substrate impurity doping must be reduced so that the drain junction can withstand higher d.c. voltages as indicated in section 536 on p/n junction breakdown voltage. The low substrate doping or high resistivity would increase the channel thickness, x_c, causing $x_c \propto$ or $> x_o$ making 3-dimensional effects important.

Figure 640.1(c) shows a wide and long channel MOSFET device structure for medium speed, low or medium power, and high-voltage switching and linear applications. The channel length is large to prevent the drain space-charge layer from punching through the channel at high drain voltages. We shall analyze this structure since it allows us to make approximations and use simple mathematics to bring out the basic physics of MOSFETs.

Modern high-speed, low-voltage ($\leq$5V) MOSFET has L< 1μm and W ~ a few microns. The theoretical minimum L for optimum performance is about 0.1 micron or 1000A. It is a three-dimensional device because the critical device dimensions (oxide thickness, x_o; channel thickness, x_c; channel length, L; channel width, W, or gate width, Z) are comparable in magnitude. The three-dimensional nature is evident in Fig.640.1(a). It shows that the electric field distribution along the channel will be influenced by a combined effect of the voltages applied to the gate, drain and source electrodes. For example, because the channel is so short, the electric field from the applied drain-to-source voltage can penetrate an appreciable fraction of the length of the short channel, resulting in the modulation of the electrical channel length by the applied drain-to-source voltage. Production micron and submicron MOSTs have a nonuniformly doped channel, $N_{AA}(y) \neq$ constant. Near the drain n+/p junction, $N_{AA}(y=L)$ is lowered to confine the drain electric field and limit its penetration into the channel. This is known as **lowly doped drain** (LDD). Composite oxide (thermal-CVD) or dual dielectric (Si_3N_4-SiO_2) are used over the LDD region, known as **spacer**, to reduce degradation due to hot electrons. In contrast, the electric field in the gate oxide of a long channel MOST, Fig.640.1(c), is influenced mainly by the voltage applied to the gate. The influence of the drain voltage is small and additive. In addition to the length effect just described, the gate width, Z, of a submicron MOST [See Fig.640.1(a)] is so small that the channel current beyond the width edges of the gate (z=0 and z=-Z) from fringing electric fields can no longer be neglected, resulting in W > Z. Consequently, the **electrical gate width**, W, enters into the MOST equation and not the **geometrical gate width**, Z. This distinction must be observed in the design of submicron and very high speed MOS transistors and integrated circuits which requires two- or three-dimensional computer-aided design (CAD) programs.

641 Conductivity Modulation Model of the MOSFET (D.C. Drift Current and Charge Control Analyses)

The conductivity modulation model of MOSFET is very simple, easy to understand, and easy to analyze. The term 'conductivity modulation' was introduced by Lilienfeld in his three 1926-1933 patents on the invention of the field-effect transistors. It was also used by Shockley and Pearson in 1948 to describe their historical first attempt to build a field-effect transistor on modern semiconductor materials such as Si and Ge. A review of the inventions and patent disclosures of the MOS and other field-effect transistors was given in 1988 by this author in an invited review article [600.1]. It seems most appropriate to introduce the MOS transistor to beginning students using the simplest electrical model, not only for historical but also pedagogical reasons owing to its simplicity in concept.

To delineate and focus on the basic principles of operation of the MOSFET with a minimum mathematical complexity, we will analyze the wide and long channel device structure, such as that shown in Fig. 640.1(c). This choice permits us to ignore the two- and three-dimensional effects. The assumptions are then: (channel width)>>(channel length)>>(oxide thickness)>>(channel thickness) or $Z>>L>>x_o>>x_c$. The four critical dimensions of a medium speed MOSFET meets essentially these inequalities. Another reason for this choice is that the two-dimensional MOSFET structure can be decomposed into two one-dimensional structures. This decomposition into one-dimension enables a clear demonstration of the physical mechanisms underlying the MOSFETs.

The ideal and simplest MOSFET is a **two-dimensional device**. The two dimensions are the directions perpendicular (x) and parallel (y) to the conduction channel. It cannot be analyzed as a purely one-dimensional device such as the MOS capacitor in chapter 4, the p/n and m/s diodes in chapter 5, or the p/n/p bipolar junction transistor in chapter 7, because the controlling electric field produced by the voltage applied to the gate is nearly transverse or perpendicular to the direction of the current flowing in the conduction channel. If the gate is narrow ($Z \simeq x_o$), the third dimension must also be taken into account due to the fringing gate electric field. Thus, the minimum model is a two-dimensional model. The two-dimensional nature is evident in Fig.610.1 or 640.1(c) for the long-channel MOST. The figures show that the controlling electric field produced by the applied gate voltage is mainly in the x-direction while electron current flowing in the channel and the electric field associated with the channel current are mainly in the y-direction.

The decomposition into two **one-dimensional** device analyses is arrived at as follows. If the gate oxide layer is thin and the channel is long, then the electric field in the oxide around the middle of the channel and over most of the channel is mainly in the x-direction and mainly controlled by the applied gate voltage. Furthermore, this x-directed oxide electric field, $\simeq V_G/x_o$, is large compared with the y-directed drift electric field in the surface channel from the applied drain-to-source voltage, $\propto V_D/L$, since the oxide is much thinner than the drain-source distance or channel length, $x_o << L$, while $V_G \propto V_D \propto 5V$. Even when the oxide field is large, it does not penetrate into the channel to affect the y-directed channel field because the oxide field is terminated by a layer of high-density semiconductor electrons (for n-channel MOSFET) in the surface inversion layer under the oxide. The charge distribution diagram of a MOS capacitor shown in Fig.412.1(b) gives N(x) which demonstrates the presence of this sheet of electron charge at the Si surface.

Indeed MOS transistors can be built that violate $x_o << L$ desired by the Dennard scaling law so that the x-directed and y-directed electric fields are not independent but interact strongly. But then the current-voltage characteristics would look like a vacuum-triode-tube whose current does not saturate like that shown in Fig.622.1(a) of a good MOSFET. Nonsaturation of the drain current is not desirable in applications. Thus, $x_o << L$ is maintained in a properly designed or scaled MOSFET whose oxide electric field is mainly in the x-direction, E_x, while its channel electric field that produces the channel and drain current is mainly in the y-direction, E_y. Then, the two fields are nearly independent and do not interact strongly since they occur in two nearly distinct regions of the MOST: E_x in the oxide layer, and E_y in the source-to-drain silicon surface channel.

This separation decomposes and simplifies the two-dimensional problem into two analytically soluble one-dimensional problems. One deals with the solution of the oxide field due to the voltage applied between the gate and the Si bulk or substrate. The other deals with the current in the channel due to the voltage applied between the drain and the source terminals.

Current-Charge Equation from Longitudinal Electric Field

With this separation of the electric fields in the x- and y-directions, we may approximate the channel current density by its y-component. This may then be integrated over the cross sectional area of the channel to give the total channel current, I_y. This must be equal to the negative of the drain current flowing into the drain contact (Kirchoff's law). The negative sign comes from our choice of current direction to make the circuit current flowing into the drain terminal positive. The drain n/p junction passes negligible current to the p-Si substrate compared with the channel current because the drain junction is reverse biased. Indeed, the drain n/p junction is reverse biased since the voltage applied to the drain is positive in order to drain out the negatively charged electrons injected from the source n/p junction through the channel towards the drain. Referring to Fig.641.1(b) for an expanded view and using the third Shockley equation, (350.3), then the drain current due to the electron drift current in the n-channel can be written as

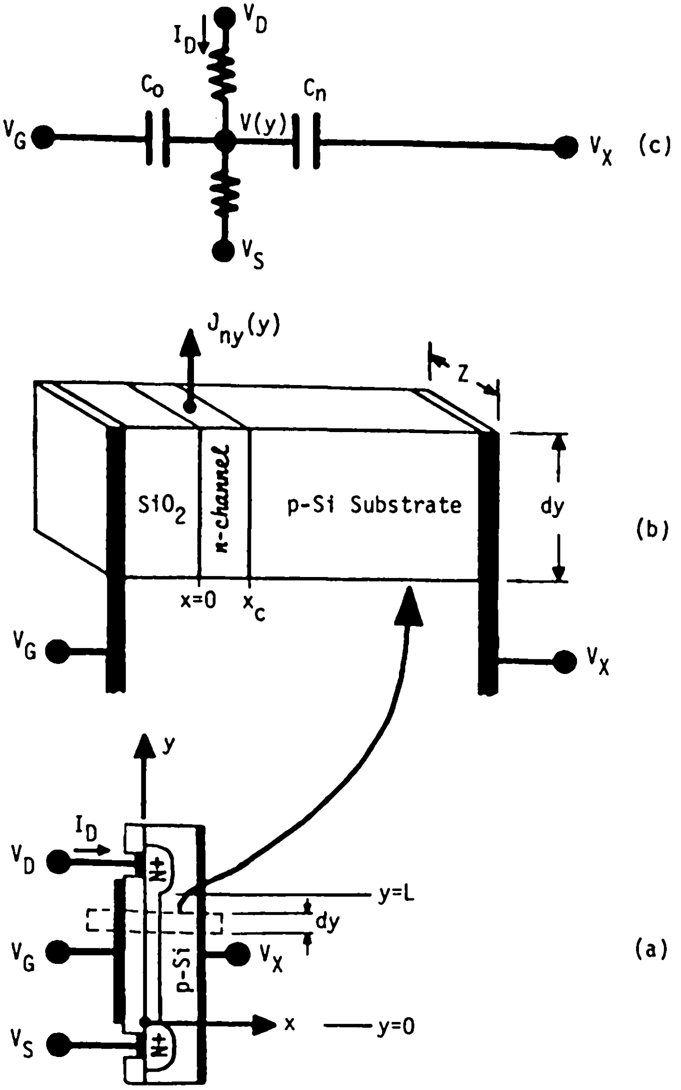

V_D
I_D
C_o
C_n
V(y)
V_G
V_X
(c)
V_S
J_ny(y)
Z
SiO_2
n-channel
p-Si Substrate
dy
(b)
x=0
x_c
V_G
V_X
y
I_D
V_D
N+
y=L
dy
V_G
V_X
p-Si
(a)
N+
V_S
x
y=0

Fig.641.1 Schematics used to analyze an n-channel MOST. (a) Cross-sectional view with x-y coordinates. (b) Enlarged view of a differential length section, dy. (c) Simplified circuit representation of the differential section.

$$- I_D = I_y = \int_0^\infty J_{ny}dxZ \tag{641.1}$$

$$\approx \int_0^\infty q\mu_n NE_y dxZ = \int_0^\infty q\mu_n N(-dV/dy)dxZ \tag{641.1A}$$

$$\approx \mu_n(-dV/dy)Z \int_0^{x_c} qN(x)dx \tag{641.1B}$$

$$\equiv \mu_n(-dV/dy)Z\cdot Q_{C'} . \tag{641.1C}$$

$Q_{C'}$ is the electron charge density (without sign) in the channel. Three assumptions are made to give (641.1B). (1) Diffusion is neglected in (641.1) so that the y-component of the electron drift current from (312.7A), $J_{ny}=q\mu_n NE_y$, and $E_y=-\partial V/\partial y$ are used to give (641.1A) while the diffusion current given in the general current equation (350.3) is ignored. (2) The electron mobility in the channel is assumed to be constant, independent of x. x is the depth distance measured from the oxide/silicon interface into the silicon substrate. The interface is located at the plane $x=0$. Thus, μ_n is moved out of the integral in (641.1A) to give (641.1B). (3) The longitudinal electric field ($E_y=-\partial V/\partial y$) is assumed to be constant across the thickness of the conduction channel, $0 < x < x_c$, so that it is also moved out of the integral in (641.1A) to give (641.1B). (2) and (3) are the key 1-dimensional approximations that decompose the 2-dimensional MOSFET into two 1-dimensional device problems.

Generally, the mobility varies with both the depth (x-axis) and length (y-axis) positions and even the width position, $\mu_n=\mu_n(x,y,z)$ due to many factors, such as the spatial (x,y,z) variations of the density of the physical imperfections, oxide/silicon interfacial atomic discontinuities, and chemical impurities, all of which scatter the electrons. The mobility and drift velocity also vary with the longitudinal (y-axis) electric field and hence the applied drain-to-source voltage. (See problem P641.1.) The electron drift velocity approaches a constant when the electric field is very high as indicated in Fig.314.1 due to electron energy loss during longitudinal optical phonon emission. In this analysis, mobility variation with the longitudinal electric field will be ignored. It is analyzed in a later section.

In the above integral, we also restricted the integration to a finite depth of x_c while the general formulae given in (641.1) extends the integration limit to ∞. This approximation is justified since few electrons pass into the p-type silicon substrate from the n+drain and n+source through the n+/p junctions and from the n+ surface inversion channel through the surface space-charge layer. More concisely, both E_y and n are negligibly small in the region $x>x_c$. A quantitatively precise definition of the channel thickness, x_c, is needed only in an advanced analysis of the

MOSFET characteristics. In this elementary drift current analysis, x_c will not even appear in the following algebra. However, its physical significance is clear - it is the thickness of the thin n-type surface channel layer which passes the electrons from the source to the drain. In this drift model, it is in fact the thickness of the strong inversion layer, x_{th}, defined for the MOS capacitor in Fig.410.1(b) or Fig.412.1(b). At very low currents when diffusion current dominates, the surface inversion is weak ($x_{th}=0$) or nonexistent ($x_I=0$). But there is still a low drain-to-source current due to diffusion, thus, x_c is generally larger than x_{th}.

Voltage-Charge Equation from Transverse Electric Field

In order to eliminate the electron charge density term in the current-charge equation (641.1C), a second relationship is required that relates the electron charge density, $Q_N=\int qN(x)dx$ in (641.1C), to the applied gate voltage, V_G. This voltage-charge relationship can be obtained by integrating the Poisson equation in the x-direction in a differential channel length of dy. It can also be obtained using the Gauss Law which is derived from integrating the Poisson equation. Figure 641.1(b) shows an expanded view of the differential section dy of the channel. Instead of integrating the Poisson equation or using the Gauss Law at each thin slice of the layer, we use the relationship between voltage and charge appearing across the MOS capacitor shown in Fig.641.1(c) which gives the same result. This capacitance and charge analysis gives

$$C_o(V_G - V) = - Q_S$$

$$= - q\int_0^\infty \rho(x)dx = - q\int_0^\infty (P - N - N_{AA})dx \tag{641.2}$$

$$= \int_0^\infty q(N - N_B)dx - \int_0^\infty q(P - P_B)dx \tag{641.2A}$$

$$= Q_N - Q_P = Q_N - Q_B. \tag{641.2B}$$

It is similar to (410.11) used to analyze the MOS capacitor where Q_N, defined by (410.22), had a negative sign which is dropped here. In (410.11), we have also included the charges trapped in the oxide traps and interface traps which are ignored here to focus on the principal charges. The induced bulk charge, Q_B in (641.2B) for an nMOST on a p-Si substrate, is defined by

$$Q_B = Q_P = - \int_0^\infty q(P_B - P)dx. \quad [Q_B<0 \text{ for nMOST}] \tag{641.2C}$$

The magnitude of the induced electron charge concentration in the n-inversion channel of the nMOST is defined by

$$Q_N = + \int_0^{\infty} q(N - N_B)dx > 0 . \qquad (641.2D)$$

The term 'induced' is used to state that the flat band condition is the reference since at flat band, we have $P(x)=P_B$=constant, and $N(x)=N_B$=constant so that $Q_P=0$ and $Q_N=0$. In contrast to (410.22) in the MOS capacitor analysis, we do not use the negative sign in Q_N here so it is also the number of electrons per unit area times the electron charge. This choice is preferred to avoid double negative sign and to make the physics easier to visualize.

Q_S is the areal density of the space charge (Coulomb/cm^2) induced in the semiconductor by the applied voltage (V_G - V) between the gate and the substrate. This semiconductor space charge is concentrated near the oxide/silicon interface. To produce an n-type surface channel in order to pass a source-drain current through the channel, the gate voltage must be positive to attract electrons to the oxide-silicon interface. The semiconductor space-charge density Q_S in (641.2) is calculated by integrating the semiconductor volume space-charge density, $\rho(x)$ (Coulomb/cm^3), over the entire thickness of the semiconductor, $0<x<\infty$. We have already used the volume space-charge density in section 242 to calculate the electron and hole concentrations in an impure semiconductor. It was zero everywhere in homogeneous semiconductor when there is no applied voltage or field. A net or non-zero volume space charge is induced in the semiconductor near the oxide-semiconductor interface by the voltage applied to the gate. It is given by $\rho(x)=q(p-n-N_{AA})$ where we assumed that the p-type semiconductor is doped with only an acceptor impurity of concentration N_{AA} and all the dopant acceptor impurity atoms are ionized.

The integral of the semiconductor space-charge density, $\rho(x)$ in (641.2), is separated into two parts in (641.2A) using $N_{AA}=P_B-N_B$ which comes from charge neutrality in regions far from the oxide/semiconductor interface. One part is due to the electrons, Q_N, and the other, holes, Q_P. The reason for this separation is that the electrons carry the channel and drain current while the holes and the negatively charged dopant acceptor impurities determine the minimum gate voltage necessary to induce the electron channel. This symmetrizes the result and also assures that the integrals of Q_N and Q_P or Q_B are finite although their integration extends over the semi-infinite space $0 \leq x \leq \infty$.

The hole charge, Q_P, gives rise to a negative charge when holes are repelled or pushed away from the semiconductor surface by a positive applied gate voltage. This repulsion leaves a layer of uncompensated and negatively charged acceptor impurity ions at the oxide/Si interface which has no holes to neutralize the negative acceptor charge. This is illustrated in Fig.410.1(b), labeled by a thickness x_d, and termed majority carrier depletion. This charge is known as the **bulk charge or bulk**

depletion layer charge. It is evident that Q_B sets the minimum (positive) gate voltage which must be applied in order to induce an n-type channel or to give a non-zero Q_N. The voltage, $-Q_B/C_o$, is usually written as $V_T=V_{GT}=-Q_B/C_o$. Note: the sign of V_{GT} is the opposite of that of Q_B and also of the neglected charges residing in the oxide, Q_{OT}, and interface, Q_{IT}, which we shall discuss in the next section, 642. V_T or V_{TH} is known as the **threshold voltage** or threshold gate voltage. It is the applied gate voltage necessary to induce a semiconductor surface inversion channel under the gate oxide and to allow an external circuit current to flow into the drain and out of the source terminals. $-Q_B/C_o$ is the minimum gate voltage necessary to turn on the MOSFET in an ideal MOSFET which has no other charge contributions to the threshold voltage, such as the oxide and interface trapped charges and the work function difference between the gate conductor and the semiconductor. In general, the bulk charge Q_B varies with both the gate and substrate voltages as well as the channel voltage V(y). The former two modify the gate threshold voltage and the latter introduces an additional drain voltage dependence in the drain current.

The above voltage-charge equations can also be obtained more systematically and quickly using the following steps. The charge neutrality condition inside the volume contained between the gate contact and bulk-substrate metal contact gives

$$Q_G + Q_S = C_oV_O + Q_S = 0 \qquad \text{from (410.28) with } Q_{IT}=Q_{OT}=0$$

or

$$C_oV_O = C_o(V_G - V) = - Q_S. \qquad (641.3)$$

The net charge contained in this volume is zero because the electric field is assumed zero inside the two metal electrodes. The charge residing on the metal or conductor electrode of the gate is C_oV_O. V_O is the voltage drop across the oxide. $C_o=\varepsilon_o/x_o$ is the oxide capacitance. It is the parallel-plate capacitance per unit area of gate oxide dielectric with a dielectric constant ε_o ($=3.8\text{x}8.854\text{x}10^{-14}$ F/cm for SiO_2) and thickness x_o. The voltage drop across the oxide or the oxide capacitance is given by $V_O=V_G-V$ as can be seen from the simple equivalent circuit shown in Fig.641.1(c). A more accurate result contains the work function difference between the metal and the semiconductor, $V_O=V_G-\phi_{MS}-V$. This was defined in (410.25)-(410.27) and derived in (412.10)-(412.12F) using the energy band of a MOS capacitor shown in Fig.421.1(c). However, we may use $V_O=V_G-\phi_{MS}-V$ as a starting equation in this elementary analysis of the drift current without going through the detailed algebra given in (412.10)-(412.12F).

642 Oxide and Interface Trapped Charges (Instability, Aging and Failure)

There are two other physical contributions to the threshold gate voltage which can be readily introduced at this beginner's level which we neglected in the preceding section. These are the trapped charges due to electrons and holes bound to traps at two locations: the oxide and interface layers. They must be added to the

MOS equations since they are the principal causes of transistor instability and integrated circuit aging and failure. Fortunately, they can be readily understood with elementary physics and chemistry and analyzed using elementary mathematics. A discussion of their effects on the MOS CV curves was given in section 403.

The first contribution is due to electrons and holes trapped at the **chemical impurities** and **physical defects** (broken bonds or dangling bonds) in the thin gate oxide layer. It is known as the **oxide trapped charge** whose areal density (Coulomb/cm^2) is denoted by the symbol Q_{OT}. The second contribution comes from the electrons and holes trapped at the chemical impurities and physical defects located at the oxide/semiconductor interfacial boundary layer. It is known as the **interface trapped charge** and its areal density (Coulomb/cm^2) is denoted by Q_{IT}.

The impurities can be residues inadvertently introduced during transistor fabrication from contamination of the gases, liquids, furnace tubes and boats, and wafer handling utensils. The physical defects (dangling silicon and oxygen bonds) could be introduced during transistor fabrication, for example, by the large thermal differential expansion of oxide film and Si at the oxide/semiconductor interface during rapid cooling of the silicon wafer at the end of a diffusion or oxidation step from the high diffusion and oxidation temperatures (about 1000C) to room temperature (22C). The presence of these traps is a cause of low manufacturing yield and high cost. Another cause is the pin holes in the oxide films.

Impurities and defects can also be introduced into the oxide film and the oxide/silicon interface during the operation of the MOS transistor. The rate of introduction increases with increasing temperature, electric field, and current density. Conditions creating high values of these three parameters are known as stress. As an example of high introduction rate, impurities **migrate** into the oxide and interface layers from a contaminated surface source, such as a dirty aluminum metal, by drift which accelerates at higher electric fields and by diffusion which accelerates at higher temperatures. As a second example, the physical defects or dangling bonds can be **generated** by **bond-breaking** or **bond-rupture** impacts from energetic electrons or holes which are accelerated by the high electric field in the oxide and near the oxide/semiconductor interface. The bonds remain broken and dangling if there is no opportunity for them to heal. This is one of the hot electron effects or hot carrier effects. Weak Si-Si and Si-O bonds may also be broken by the capture of holes which releases a sufficient amount of recombination energy to break the bonds. These are the two principal causes of transistor instability which cause **aging** and **failure** and reduce the **reliability** of integrated circuit chips. They are the two most important subjects in transistor reliability physics research.

Both the yield and reliability are increasingly important obstacles to the manufacturing of higher performance integrated circuits since higher performance is obtained by reducing the transistor dimensions such as the oxide thickness and the source-drain separation or channel length. Smaller dimensions make the submicron transistor more susceptible to smaller defects. Smaller dimension also results in higher electric fields (across the oxide layer and between the source and drain) since the power supply voltage cannot be reduced from the nominal 5V by more than a factor of about three to four (down to 1.5V battery level) to avoid physical and circuit noise. Higher fields accelerate the aging and failure mechanisms. These practical limitations on performance, yield, and operating life, have attracted increasing concern and dedicated focus in the research and development of submicron silicon transistors and integrated circuits.

In the earlier literature and textbooks as well as some recent textbooks, researchers and authors have not recognized the importance of the spatial separation and the different origin of the oxide and interface charges or traps. Thus, these traps are lumped into one term, known as the **surface states**, **surface traps** or surface trapped charges, and denoted by the symbol Q_{SS} or Q_{ST} where subscripts SS stand for surface states and ST for surface traps. While reading these research and engineering literature and textbooks, the following replacement is to be made and the physical origin of Q_{OT} and Q_{IT} just discussed are to be noted.

$$Q_{SS} \equiv Q_{ST} \equiv Q_{OT} + Q_{IT}. \tag{642.1}$$

All three terms above have an associated sign, negative for electron-like charge and positive for hole-like charge.

Contributions to the threshold voltage from the oxide and interface charges can be readily included in the foregoing mathematical analysis by not setting Q_{OT} and Q_{IT} to zero in (410.28) which gave (641.3). These trapped charges were included in the analysis of the C-V curves of the 1-d MOS capacitor in chapter 4. The MOSC results can be used for MOST here since the 2-d MOST is decomposed into two 1-d problems: along the x-axis (this subsection) and the y-axis (preceding subsection), and since the x-axis MOST problem is identical to the MOS capacitor problem analyzed in chapter 4 (which also used the x-axis). The only difference is the addition of a nonequilibrium concentration of inversion electrons in the MOST which are supplied by the n+ source and extracted by the n+ drain of the MOST. The following results from MOSC analysis can also be derived by integrating the Poisson equation or using the Gauss Law. The modified (641.2B) and (641.3) are

$$C_oV_O = C_o(V_G - \phi_{MS} - V) = Q_N - (Q_B + Q_{OT} + Q_{IT})$$

so

$$Q_N = C_o(V_G - V_{GT} - V). \tag{642.2}$$

Here

$$V_{GT} = \phi_{MS} - (Q_B + Q_{OT} + Q_{IT})/C_o \tag{642.3}$$

is the threshold voltage. It contains four contributions: the bulk charge, the charged oxide traps, the charged interface traps, and the metal-semiconductor work function difference.

One of the key ideas that gives the simple result (642.3) was discussed in chapter 4 on MOS capacitors concerning the distributed oxide charge whose volume density is $\rho_{OT}(x)$ as shown in Fig.412.1(b). To recapitulate, Q_{OT} is the effective charge sheet located at the oxide/silicon interface, $x=0$, which would produce the same amount of deflection of semiconductor energy band at $x=0$ as the oxide charge distributed through the entire oxide. Thus, Q_{OT} is not the integrated volume density of the oxide charge distribution, $\int \rho_{OT}(x)dx$, but is the first moment which can be easily derived using the parallel capacitance model. It is given by

$$Q_{OT} = \int_0^{x_o} (x/x_o)\rho_{OT}(x)dx. \qquad (642.4)$$

The oxide electric field would no longer be constant as depicted in Fig.412.1(c) but highly distorted with maxima and minima determined by the sign and the magnitude of the volume density of the charges trapped in the oxide layer. The field and potential variations can be obtained only by solving the Poisson equation through the oxide layer.

643 The MOSFET Equations and D.C. Characteristics

We now have two equations, current-vs-charge and voltage-vs-charge, so we can eliminate the charge Q_N between (641.1C) and (642.2). These are repeated below for ease of reference and to show the approximation of the inversion layer charge by the induced electron charge.

$$-I_D = I_y = \mu_n(-dV/dy)Z\int_0^{x_C} qNdx = \mu_n(-dV/dy)ZQ_C \qquad (641.1A)$$

$$Q_{C'} = \int_0^{x_C} qN(x)dx \qquad (642.2x)$$

$$\approx \int_0^{\infty} q[N(x)-N_B]dx = Q_N \qquad (642.2y)$$

$$= C_o(V_G - V_{GT} - V). \qquad (642.2)$$

$Q_{C'}$ is the electron charge density (C/cm^2) in the inversion channel while Q_N is the induced electron charge density (C/cm^2) using flat band as the reference as defined in (642.2y) which had a negative sign in (410.22) that is dropped here. The approximation of $Q_{C'}$ by Q_N, shown by $\approx$ in (642.2y) above, introduces very little error in the drift current model since the 'leakage' current contributed by the equilibrium concentration of minority carrier, N_B, is negligible compared with the large drift current. At low currents when diffusion dominates, this approximation will become increasingly inaccurate and this 'leakage' or diffusion current plus the drain junction leakage current must be added.

Eliminating $Q_{C'}$ in (641.1A) using the Q_N approximation given in (642.2), the fundamental differential equation of the MOSFET in one dimension is

$$\boxed{I_D = \mu_n C_o Z(V_G - V_{GT} - V)(dV/dy).} \qquad (643.1)$$

This is the basic differential equation of the long-channel MOSFET. We repeat the meaning of the parameters of this equation in the following paragraphs.

I_D is the external circuit current flowing into the drain terminal or drain electrode. It is equal to the negative of the channel current, I_Y. The negative sign comes from the positive direction of the y-axis in the coordinate system used in Figs.641.1(a) or (b) and positive direction of the circuit current referenced as that flowing into the device terminal or device electrode. The coordinate axis is selected here to make the physical picture of the MOS transistor oriented exactly the same as its circuit symbol and as its equivalent circuit. This choice avoids mental reorientation when considering a device-physics or technology-chemistry phenomenon during a circuit analysis of MOSFET operation.

μ_n is the average mobility of the electrons in the n-type semiconductor surface inversion channel under the gated oxide-semiconductor interface on the p-type substrate. It is known as the **conductivity mobility**. Under our simplified assumption, it is also equal to the effective mobility, transfer mobility, or transconductance mobility. But in general, the effective mobility is not equal to the conductivity mobility. Then, conductivity mobility is the correct one to use in this equation. The value of the conductivity mobility is smaller than the bulk mobility at the same bulk dopant impurity concentration because there are additional scatterings by chemical impurities and physical defects at the SiO_2/Si interface. At very high channel or drain current densities, which occurs in very thin channels and at high applied gate voltages, the electron motion is confined in the y-z plane. The x-direction is quantized due to the presence of the narrow and deep surface potential well. The concentration and mobility of the electrons in the surface channel are modified by quantization in this 1-dimensional potential well.

E_y or $-dV/dy$ is the electric field along the surface channel which is directed along the y-axis.

Z is the geometrical width of the gate. We assume a wide gate structure. Thus, the electrical channel width W_E is equal to the geometrical gate width, Z. Normally, $W_E > Z$ due to fringe electric field at the edges or perimeter of the gate electrode. The wide gate structure assumes that Z is much larger than the oxide thickness x_o, the channel thickness x_c, and the channel length L, so that the fringe electric field does not widen the channel significantly.

x_c is the thickness of the conduction channel which contains all the conduction electrons in the thin channel layer. It is assumed to be much smaller than the gate oxide thickness, the channel width, and the channel length, so that the electron density can be approximated by a delta function concentrated at the oxide/semiconductor interface. Finite thickness would modify the characteristics which can be predicted in an advanced analysis of the MOSFET. The finite channel thickness and its effects do not appear in the elementary MOS differential equation given by (643.1) since the electron charge density in the inversion channel $0<x<x_c$, $Q_{C'}$, is approximated by the total induced electron charge density referenced to the flat-band condition, Q_N.

C_o is the oxide capacitance per unit area given by $C_o=\varepsilon_o/x_o$ in F/cm^2. It is the parallel-plate capacitance of a dielectric layer of thickness x_o and dielectric constant ε_o. For SiO_2, $\varepsilon_o=3.8x8.854x10^{-14}$ F/cm. $8.85x10^{-14}$ F/cm is the permittivity of free space. For a 1000A (100nm) oxide, $C_o=3.8x8.85x10^{-14}/10^{-5}=3.36x10^4$ pF/cm^2. For a state-of-the-art submicron Si MOSFET, x_o is ten times thinner, 100A (10nm), and C_o is ten times higher. For a 1μm^2 gate area, $C_oA_g=3.36x10^5x10^{-8}=3.36$fF.

V_G is the d.c. voltage applied to the gate electrode using the substrate and the source as the common reference, i.e., $V_X=V_S=0$. In circuit applications described later, $V_X\neq 0$ and $V_S\neq 0$ are encountered and the MOSFET equation is modified appropriately.

$V(y)$ is the potential (or voltage if stated very loosely) at the oxide/semiconductor interface in the channel at a distance y from the source end of the channel. In an advanced analysis, V(y) is the quasi-Fermi potential of the minority carriers, electrons in the n-type inversion channel MOS transistor, $V_N(y)$. The quasi-Fermi potential of the majority carriers, holes in this case, would be constant given by the bulk value of the p-type substrate and chosen as the reference voltage, $V_P(x)=V_P=\text{constant}=0$. If the substrate is biased, $V_X\neq 0$, then $V_P=V_X$.

V_{GT} is the threshold gate voltage measured relative to the silicon substrate. It is the minimum gate voltage required to produce an n-type surface channel on a p-type Si substrate in order to cause a current to flow between the source and the drain. It contains six contributions from two imperfections, the chemical impurities and the physical defects, at three locations : (i) the impurity ions in the surface space-charge layer of the semiconductor, Q_B, (ii) the impurity ions at the oxide/semiconductor interface, Q_{IT}, (iii) the impurity ions in the oxide layer, Q_{OT}, (iv) the charged physical defects or dangling bonds in the semiconductor surface space-charge layer underneath the oxide, (v) the charged physical defects or dangling bonds at the oxide/semiconductor interface, Q_{IT}, and (vi) the charged physical defects or dangling bonds in the oxide layer, Q_{OT}. The seventh component, (vii), metal/semiconductor work function difference, ϕ_{MS}, is added later using the E-x energy band.

The solution of the long-channel MOSFET differential equation, (643.1), can be obtained by integration from the source at $y=0$ to the drain at $y=L$ since at these two points, the voltage $V(0)$ and $V(L)$ are known: $V(0)=V_S$ and $V(L)=V_D$. Assuming that the oxide, interface and bulk charges are spatially constant, the integration then gives

$$\int_0^L I_D dy = \int_{V_S}^{V_D} Z\mu_n Q_N(V)dV = \int_{V_S}^{V_D} Z\mu_n C_o(V_G - V_{GT} - V)dV$$

or

$$I_D L = Z\mu_n C_o[(V_G-V_{GT})(V_D-V_S) - ½(V_D^2 - V_S^2)]$$

$$\boxed{I_D = (Z/L)\mu_n C_o[(V_G-V_{GT})(V_D-V_S) - ½(V_D^2 - V_S^2)].} \qquad (643.2)$$

This is the long-channel 4-terminal d.c. MOSFET equation with the semiconductor substrate grounded or used as the reference voltage electrode. If the substrate or bulk silicon is biased at an applied voltage of V_X relative to a common ground reference voltage, then V_G must be replaced by V_G-V_X. The bulk charge term, Q_B, would also be modified since a reverse bias applied to the substrate (V_X is negative on p-Si substrate relative to the n-type diffused source and drain regions) would significantly increase the thickness of the space-charge layer. This gives larger negative bulk charge at the source junction which would increase the threshold voltage. This substrate bias effect on the threshold voltage is known as the **body effect**. It made a very significant impact on the application history of MOSFET in the early 1970's. It started Intel into volume production of the first 4k-bit dynamic random access memory (DRAM) chips, the Intel-2104, in 1972 and enabled IBM to replace the magnetic core memory in mainframe computers by Si MOS dynamic random access memory (DRAM) and to ship in volume an all-semiconductor (Si) mainframe computer, the IBM model 370/158 in 1973 [600.1]. During this early generation of silicon integrated circuit technology, it was necessary to use a d.c. substrate bias to compensate the residual positive oxide charge in order to cut off the built-in n-channel at zero applied gate voltage. Current technology uses an implanted low-concentration layer of boron ions in the silicon surface layer beneath the gated oxide. The implanted boron acceptors provide the additional negative charge (Q_B is increased) to compensate for the residual positive oxide charge. It also compensates for the negative gate voltage shift due to a gate-conductor/Si work function difference.

A simplified solution is obtained when both the source and the substrate are grounded and serve as the common terminal for the input-gate and the output-drain. For this three-terminal case, the long-channel d.c. MOSFET equation, (653.2), simplifies to

$$I_D = (Z/L)\mu_n C_o[(V_G-V_{GT})V_D - \tfrac{1}{2}V_D^2]. \qquad (643.3)$$

A family of output characteristics, I_D versus V_D, with the gate voltage, V_G-V_{GT}, as the parameter is shown in Fig.643.1(a). A family of transfer characteristics, I_D versus V_G-V_{GT}, with the drain voltage V_D as the parameter is shown in Fig.643.1(b). I_D becomes a constant when V_D reaches V_G-V_{GT}. This is known as drain current saturation. At higher drain voltage, $V_D > V_G$-V_{GT}, the drain voltage has no control on the channel current since electron density at the drain junction end in the channel has dropped to zero, i.e. $Q_N(y=L)=0$.

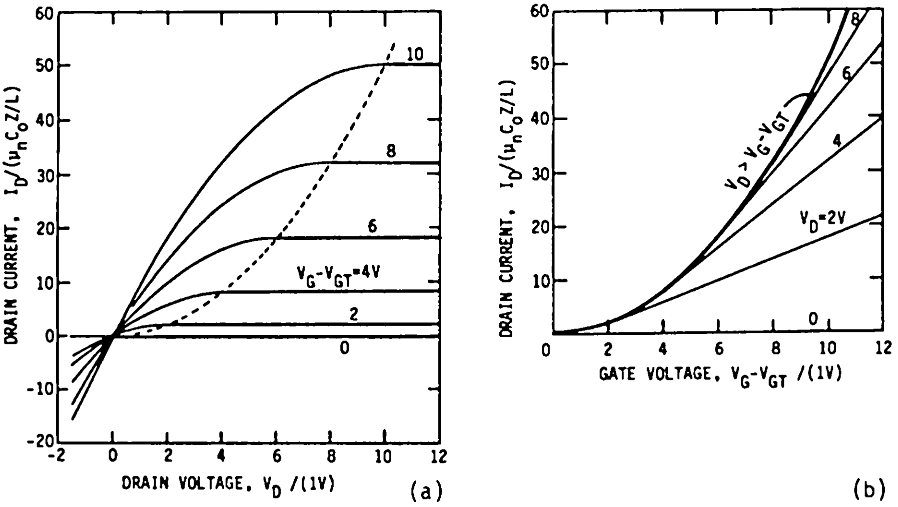


Fig.643.1 Families of normalized D.C. characteristics of an n-channel MOSFET. (a) Output or drain characteristics, I_D vs V_D, with the gate voltage, V_G-V_{GT}, as the parameter. (b) Transfer characteristics, I_D vs V_G-V_{GT}, with the drain voltage, V_D, as the parameter. Note: Normalization current, $(\mu_n C_o Z/L)$, is in unit of volt^2.

The drain conductance of MOS transistor can be derived by taking the derivative of (643.3) and is given by

$$g_d = (\partial I_D/\partial V_D)\Big|_{V_G} = (Z/L)\mu_n C_o[V_G - V_{GT} - V_D]. \qquad (643.4)$$

It decreases linearly with increasing drain voltage, V_D. At $V_D = V_G - V_{GT}$, the drain conductance and the electron charge, $Q_N(y=L)$ are both zero. Beyond this point, the drain conductance is zero and the drain current is a constant which is no longer controlled by the drain voltage. It is only affected by the gate voltage. The drain voltage at $g_d=0$ and $Q_N(y=L)=0$ is known as the **current-saturation drain voltage** and denoted by the symbol, V_{DS}:

$$V_{DS} = V_G - V_{GT}. \qquad (643.5)$$

The drain saturation current is then given by the **parabolic equation**

$$I_{DS} = I_D(\text{at } V_D = V_{DS} = V_G - V_{GT}) = (Z/L)(\mu_n C_o/2)(V_G - V_{GT})^2. \qquad (643.6)$$

This parabola is the thick and darker asymptotic curve in the drain transfer characteristics given in Fig.643.1(b). The parabolic relationship has been used to advantage in linear push-pull stereo amplifier applications to cancel out the second harmonic distortion in the load from the two push-pull MOSFETs.

The drain current equations given by (643.2) or (643.3) are no longer valid when $V_D > V_G - V_{GT}$. If they are used regardless, the computed drain current would decrease with increasing drain voltage, giving a physically unrealistic negative resistance. When Shockley invented the first successful field-effect transistor, the junction-gate FET, in 1952 [600.1], this invalid behavior was recognized. A detailed explanation of the current saturation physics follows.

When the drain voltage is greater than $V_G - V_{GT}$, the solution given by (643.3) is no longer valid because Q_N at $y \geq L$ becomes negative which gives a negative electron number. Thus, at $V_D = V_G - V_{GT}$ and when $V_D > V_G - V_{GT}$, we have $Q_N(y \geq L) = 0$ and not negative Q_N, which is the **carrier depletion** condition and was used previously in p/n junction analysis in chapter 5 and MOSC analysis in chapter 4. At this point, the drain voltage loses its influence on the number of electrons that can be injected by the source. This is due to the presence of the depletion layer, where $Q_N = 0$, at the drain end of the channel which prevents the drain electric field from penetrating into the channel towards the source to pull out more electrons from the source into the channel. Thus, the channel current will be completely controlled and can only be controlled by the gate voltage and the transverse oxide electric field at the source, $y=0$, which can still lower the potential barrier of the n+/p source junction to 'inject' more electrons from the n+ source into the n-type surface inversion layer.

The constancy of drain current is not maintained in short-channel MOSFETs since the additional drain voltage beyond V_G-V_{GT} will cause the $Q_N(y_L)=0$ point to move slightly into the channel towards the source in order to deplete more electrons. In short channels, this slight penetration may become a significant fraction of the channel length. Then, the electrical length of the channel is shortened by the excess drain voltage, $V_D - V_{DS} = V_D - (V_G - V_{GT})$. The shortened electrical channel would cause the source to inject more electrons into the channel and the channel current would increase in inverse proportion to the electrical channel length. This also gives a finite and small drain conductance instead of the zero drain saturation conductance predicted by the ideal long-channel MOSFET equation, (643.4). The nonsaturating drain current and finite, non-zero drain conductance are two of the **channel shortening** or **channel length modulation effects** which are commonly known as the **short channel effects**. Another short channel effect is associated with short geometrical length which lowers the threshold voltage because the drain voltage, which has the same sign as the gate voltage, will lower the threshold gate voltage. An elementary mathematical analysis of the channel shortening at high drain voltage can be made using the depletion layer theory of the reverse biased p/n junction diode described in section 531.

Another important parameter of the MOSFET is the **transconductance** defined by

$$g_m = (\partial I_D/\partial V_G)\Big|_{V_D} = (Z/L)\mu_n C_o V_D. \tag{643.7}$$

This also defines the **effective mobility** or the **transconductance mobility**. In the drain current saturation range, it becomes

$$g_{ms} = (Z/L)\mu_n C_o V_{DS} = (Z/L)\mu_n C_o (V_G - V_{GT}). \tag{643.8}$$

This shows that the transconductance increases linearly with the drain voltage and becomes a constant when the drain voltage is equal to or greater than the drain saturation voltage, V_{DS} $(=V_G-V_{GT})$. The saturation transconductance is also related to the drain saturation current given by (643.6). It is

$$g_{ms} = \sqrt{(2L/Z\mu_n C_o)I_{DS}} \tag{643.9}$$

showing a square root dependence on the drain current.

A third differential parameter of the MOSFET is the **voltage amplification factor** defined by

$$\mu = -(\partial V_D/\partial V_G)\Big|_{I_D} = V_D/(V_G - V_{GT} - V_D). \tag{643.10}$$

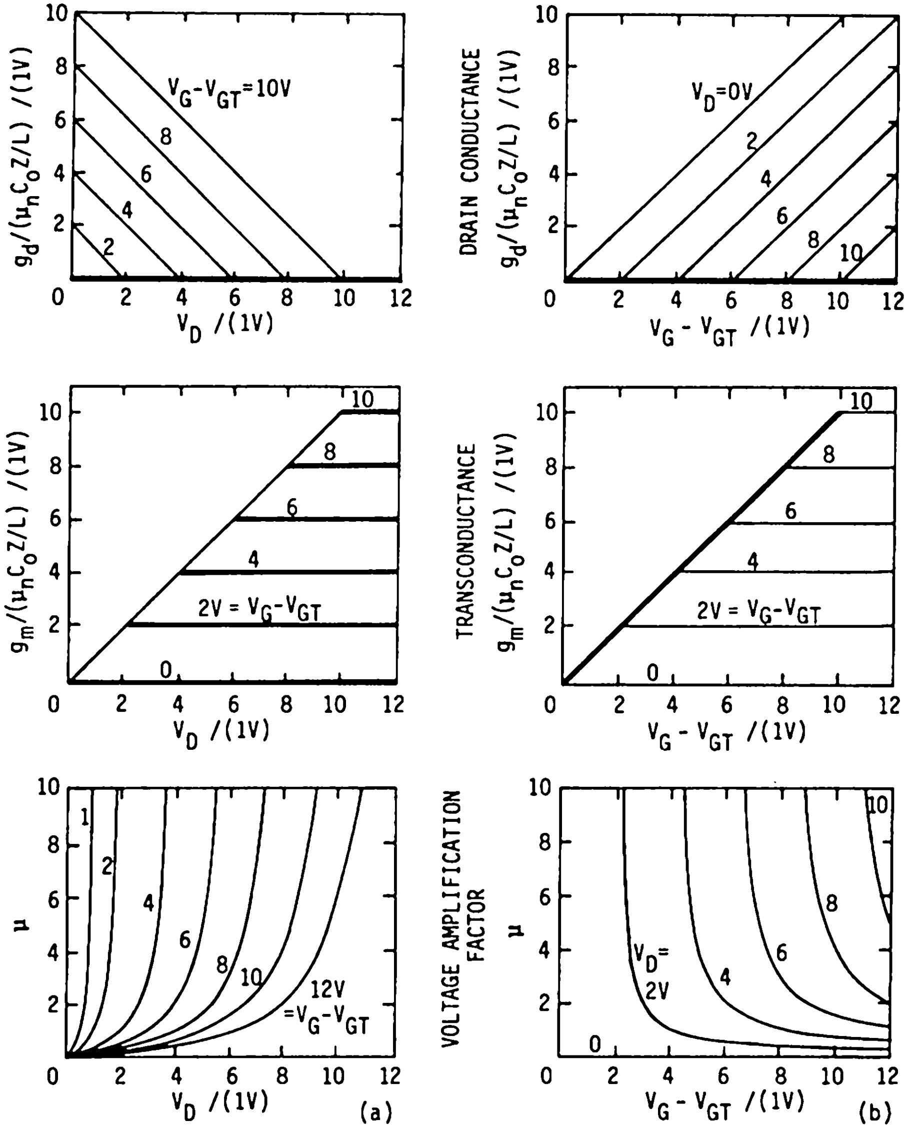


Fig.643.2 Drain conductance, transconductance and voltage amplification factor of an nMOST (a) as a function of drain voltage at a range of gate voltages and (b) as a function of gate voltage at a range of drain voltages.

The voltage amplification factor just given, (643.10), diverges at drain current saturation when $V_D = V_{DS} = V_G - V_{GT}$. This is physically not possible without an internal or external positive feedback path. The false result is caused by the long-channel approximation. It is also directly related to the zero drain conductance at saturation since

$$\mu = g_m/g_d. \tag{643.11}$$

Thus, when the finite drain conductance at and beyond saturation due to drain field penetration into the channel is taken into account, the divergence of the voltage amplification factor at and beyond saturation is also removed.

A family of normalized curves of g_d, g_m, and μ are given as a function of V_D and V_G in Figs.643.2(a) and (b) on the previous page.

644 Numerical Examples of MOSFET D.C. Characteristics

A numerical example is given to illustrate the orders of magnitude of the MOSFET parameters and characteristics. Consider an n-channel MOSFET whose oxide thickness is 1000A and whose electron mobility in the n-type surface channel is 600 cm^2/V-s. Then, the oxide capacitance per unit gate area is

$$\begin{aligned} C_o = \varepsilon_o/x_o &= 3.8\text{x}8.85\text{x}10^{-14}/10^{-5} = 3.36\text{x}10^{-8} \quad \text{F/cm}^2 \\ &= 3.36\text{x}10^{4} \quad \text{pF/cm}^2 \\ &= 0.336 \quad \text{fF/}\mu\text{m}^2. \end{aligned}$$

The saturation drain current and transconductance are respectively

$$I_{DS} = (Z/L)\mu_n C_o (V_G - V_{GT})^2/2 = 10(Z/L)(V_G - V_{GT})^2 \ \mu\text{A}$$

and

$$g_{ms} = (Z/L)\mu_n C_o (V_G - V_{GT}) = 20(Z/L)(V_G - V_{GT}) \ \mu\text{S}.$$

The numerical values of a wide gate and long channel MOSFET with $Z = 20\mu$m and $L = 5\mu$m at $V_G - V_{GT} = 5.0$V are given as follows. The gate capacitance is

$$C_o A = C_o ZL = 0.0336 \text{ pF} = 33.6 \text{ fF}.$$

which is quite small and hence capacitance from parasitics and interconnect lines must be included. The value, 33.6fF, if charged up to 5V, gives $5\text{x}33.6\text{x}10^{-15}/1.60\text{x}10^{-19} = 10^5$ electrons which is bordering the minimum to avoid noise. The saturation drain current is

$$I_{DS} = 40(V_G - V_{GT})^2 \ \mu\text{A} = 1000\mu\text{A} = 1\text{mA}.$$

Using 16,000 of this nMOST in one 16k DRAM chip would have required 16A current if all 16,000 MOSTs were turned on simultaneously. However, only one cell or at most eight cells (eight bits or one byte) are accessed simultaneously. Much of the power dissipation comes from the nMOST line drivers which must have large aspect (Z/L) ratio to drive the large capacitance of the line. A typical 16k DRAM is rated at a few hundred mA and 5V to meet the heat dissipation limitation of the dual-in-line 22-pin or 18-pin package. The MOST in the memory cell has an aspect ratio Z/L of about 2 instead of the assumed 20/5=4 and $(V_G-V_{GT})=4V$. These lower the power dissipation further.

The transconductance of this nMOST is

$$g_{ms} = 80(V_G-V_{GT})\ \mu S = 400\mu S.$$

at $V_G-V_{GT}=5V$. In comparison, the transconductance of a bipolar junction transistor (BJT), to be described in chapter 7, is $g_m = qI/kT = 1mA/25mV = 0.04S = 40mS$ at 1mA. It is 100-times larger than this nMOST at the same current. **The low transconductance is a drawback of all FETs not just MOSFET.** Thus, FETs have a much lower capability to drive the large capacitance load from a long conductor line which is necessary to interconnect many transistors. A large capacitance load will slow down a FET circuit very substantially.

The calculation can be extended to a state-of-the-art submicron MOST in an Si VLSI chip. For a submicron MOS chip, we assume $x_o=100A$ and $Z=L=1\ \mu m$. Then, $C_o=3.36fF$, $I_{DS}=2500\mu A$, and $g_{ms}=1000\mu S=1mS$. For a 256k bit DRAM chip of about 100x200 mil^2 (0.129cm^2) area, the total current would be 640A and the total power, 3200 watts, with a power density, 24800 W/cm^2. If all the MOST's were turned on simultaneously the chip would be destroyed by heating. During read, only one cell is accessed, requiring 2.5mA or 12.5mW. Most of the rated 500-1000mW power dissipation comes from the sense-amplifier/line-driver transistors which have large Z/L in order to drive the load capacitances.

650 SMALL-SIGNAL EQUIVALENT CIRCUIT MODEL OF MOSFET

The drain conductance and the transconductance of a MOSFET were already obtained from the d.c. drain current equation on the preceding pages. These quantities were low frequency values. Their derivation did not take into account the distributed nature of the signal propagation through the channel of the MOSFET. That is, the channel was treated as a single lumped circuit element rather than a transmission line. In this introductory text, we will use the lumped model to compute the capacitance elements of the small-signal equivalent circuit model. The single-lump circuit model approach is also known as the **charge control analysis** or the stored charge analysis since the stored charge is the principal quantity used in the analysis. For example, the channel current was related to the electron charge stored in the channel as indicated by (641.1A) and (642.2) whose approximations were described in section 643.

$$I_D \approx \mu_n Z(dV/dy)Q_N \tag{641.1A}$$

$$Q_N = \int_0^{\infty} q[N(x)-N_B]dx = C_o[V_G - V_{GT} - V(y)]. \tag{642.2}$$

Equation (641.1A) was then integrated over the length of the channel using $Q_N(y)$ of (642.2) to give the terminal I_D-V_D equations, (643.2) for a biased source, $V_S \neq 0$, or (643.3) for grounded source, $V_S=0$. These can be written in slightly different and more convenient forms to obtain the stored charge in each region of the transistor.

$$I_D = (Z/L)(\mu_n C_o/2)[(V_G-V_{GT}-V_S)^2 - (V_G-V_{GT}-V_D)^2] \tag{650.1}$$

and for $V_S=0$,

$$I_D = (Z/L)(\mu_n C_o/2)[(V_G-V_{GT})^2 \quad - (V_G-V_{GT}-V_D)^2]. \tag{650.2}$$

Then, the channel electric field at a distance y from the source at $y=0$ can be obtained from (641.1A) and (642.2) using I_D in (650.2). It is given by (for $V_S=0$)

$$\begin{aligned} dV/dy &= I_D/(\mu_n Z Q_N) \\ &= \frac{[V_G-V_{GT}]^2-[V_G-V_{GT}-V_D]^2}{2L[V_G-V_{GT}-V(y)]}. \end{aligned} \tag{650.3}$$

This is used to change the integration variable from y to V in order to obtain the stored charge in the various regions of the transistor. In the following sections, the capacitance elements as well as the cut-off frequencies are obtained using the charge control analysis.

651 Charge Control Analysis of Capacitance Elements

The capacitances of the **intrinsic transistor** can be computed by first finding the expressions of the charges stored in each region or on each electrode. In the conventional four terminal circuit analysis, the capacitance looking into two terminals (gate-source or drain-source) of the MOST is then given by $\partial Q/\partial V$. The other terminal voltage is kept constant to give the short-circuit capacitance while the other terminal current is kept constant to give the open-circuit capacitance. The intrinsic transistor is defined as the portion of a real transistor that excludes any parasitic capacitances, inductances and resistances. The charge stored in the channel of width W and length L, Q_C, is obtained using (650.2), (650.3) and (641.2D). The detailed algebra is given as follows.

$$Q_C = +\int_0^{\infty}\int_0^{L} q(N-N_B)dxdyZ = +Z\int_0^{L} Q_N(y)dy = +Z\int_0^{V_D} Q_N(V)(dy/dV)dV$$

$$= +(\mu_n Z^2/I_D)\int_0^{V_D} Q_N^2 dV \tag{651.1}$$

$$= +2C_oLZ\,\frac{\displaystyle\int_0^{V_D}(V_G-V_{GT}-V)^2dV}{(V_G-V_{GT})^2-(V_G-V_{GT}-V_D)^2} \tag{651.2}$$

$$= \frac{2}{3}C_oLZ\left[\frac{(V_G-V_{GT})^3-(V_G-V_{GT}-V_D)^3}{(V_G-V_{GT})^2-(V_G-V_{GT}-V_D)^2}\right] \tag{651.3}$$

$$= C_oLZ\left[+(V_G-V_{GT})-\frac{3(V_G-V_{GT})V_D-2V_D^2}{3[2(V_G-V_{GT})-V_D]}\right]. \tag{651.4}$$

The charge stored on the gate can also be obtained by a similar procedure such as integrating $C_oV_O(y)$ from $y=0$ to $y=L$. However, a short-cut can be used which eliminates much of the algebra by noting that the Gauss Law was used for the gate charge and the channel charge. From (642.2), $C_oV_O(y) = -Q_{SS}-Q_B+Q_N(y) = C_oV_{GT}+Q_N(y)$. Thus,

$$Q_G = \int_0^{L} C_oV_o(y)dyZ = \int_0^{L}[C_oV_{GT}+Q_N(y)]dyZ \tag{651.5}$$

$$= C_oLZV_{GT} + \int_0^{V_D} Q_N(V)(dy/dV)ZdV = C_oLZV_{GT} + Q_C \tag{651.6}$$

$$= C_oLZ\left[V_G-\frac{3(V_G-V_{GT})V_D-2V_D^2}{3[2(V_G-V_{GT})-V_D]}\right]. \tag{651.7}$$

Since $Q_G = C_oLZV_{GT} + Q_C$, or Q_G and Q_C differ by a constant C_oLZV_{GT}, only the gate charge, Q_G, is computed as a function of V_D and V_G. The families of Q_G as a function of V_D or V_G with the other as the constant parameter are plotted respectively in Figs.651.1(a) and (b).

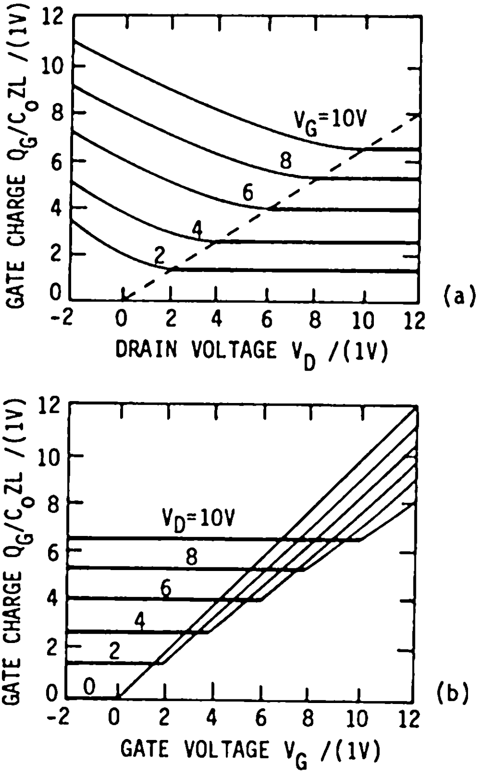


Fig.651.1 The stored charge on the gate electrode as a function of (a) V_D and (b) V_G.

The short-circuit gate and drain capacitances are then given by

$$C_{GS} = (\partial Q_G/\partial V_G)\Big|_{V_D} = C_oLZ\left[1 - \frac{V_D^2}{3[2(V_G-V_{GT})-V_D]^2}\right]$$

$$= C_oLZ[\ 1 - (V_D/V_{DS})^4(I_{DS}/I_D)^2/3\ ] \qquad (651.8)$$

$$C_{DS} = (\partial Q_C/\partial V_D)\Big|_{V_G} = \frac{2}{3}C_oLZ\left[1 - \frac{(V_G-V_{GT})^2}{[2(V_G-V_{GT})-V_D]^2}\right]$$

$$= \frac{2}{3}C_oLZ[\ 1 - (V_D/V_{DS})^2(I_{DS}/I_D)^2\ ]. \qquad (651.9)$$

The open-circuit gate capacitance is then given by

$$C_{GO} = (\partial Q_G/\partial V_G)\Big|_{I_D}$$

$$= C_oZL\left[1 - \frac{V_D{}^2 + 2[3(V_G-V_{GT}) - V_D](V_G-V_{GT}-V_D)(\partial V_D/\partial V_G)}{3[2(V_G-V_{GT}) - V_D]}\right]$$

$$= C_oLZ2(V_G-V_{GT})/[2(V_G-V_{GT}) - V_D]. \qquad (651.10)$$

The open-circuit drain capacitance, defined by $C_{DO}=(\partial Q_C/\partial V_D)$ at I_G=constant, cannot be computed by the charge analysis method we have used so far since there is no d.c. gate current because the gate dielectric is assumed to be a perfect insulator. It can be computed if we assume that there is also no transient gate current, i.e. $I_G=dq_G/dt=0$ or Q_G=constant. This will be left as an exercise. Instead, we will compute it from an assumed small-signal equivalent circuit model of the MOSFET. Small-signal equivalent circuit models are not unique, that is, one can draw several circuits which would all give the same terminal impedances or admittances such as the input and output admittances, the transfer admittance, and the voltage gain. However, one requirement on the circuit elements of the equivalent circuit is that they must be independent of the signal frequency. This frequency independence imposes a restriction on the choice among the equivalent circuit models. One of the small-signal equivalent circuit models for the MOSFET is the pi (π) model shown in Fig.651.2. From this model, we can obtain the open-circuit input admittance which is

$$y_{gO} = j\omega C_{gs} + \frac{j\omega C_{gd}(g_m + g_d) - \omega^2 C_{ds}C_{gd}}{g_d + j\omega(C_{ds} + C_{gd})}. \qquad (651.11)$$

The low frequency open-circuit gate capacitance can be obtained from y_{gO} by letting $\omega\to 0$ in (651.11) so that the ω^2 term drops out and y_{gO} becomes purely imaginary or capacitive. The result is

$$j\omega C_{GO} = \lim_{\omega\to 0} y_{gO} = j\omega C_{gs} + j\omega C_{gd}(g_m + g_d)/g_d \qquad (651.12)$$

giving

$$C_{GO} = C_{gs} + C_{gd}(g_m + g_d)/g_d. \qquad (651.13)$$

The short-circuit capacitances from the equivalent circuit model are

$$C_{GS} = C_{gs} + C_{gd} \qquad (651.14)$$

and

$$C_{DS} = C_{ds} + C_{gd}. \qquad (651.15)$$

C_{gd} is known as the **feed-through** capacitance which couples the input directly to the output and causes undesirable oscillation.

These expressions of C_{GO}, C_{GS} and C_{DS} from the equivalent circuit model are used to obtain the analytical expressions of the small-signal equivalent circuit elements by equating them to their charge control expressions given by (651.8), (651.9) and (651.10). Note that there are three unknown capacitance elements, C_{gs}, C_{gd} and C_{ds}, whose d.c. voltage dependences are needed and we have just three expressions, C_{GO}, C_{GS} and C_{DS}, to derive a set of unique solutions for the three unknowns.

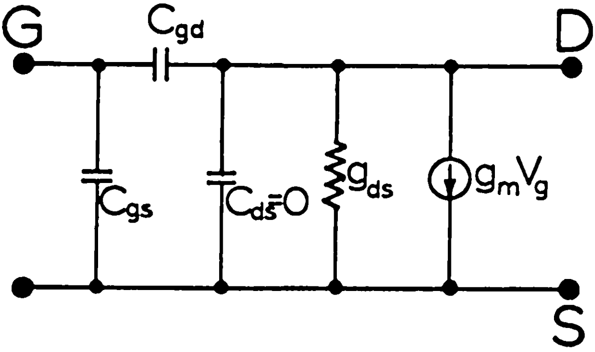


Fig.651.2 The small-signal hybrid-pi equivalent circuit model of a MOSFET.

C_{gd} can be obtained by eliminating C_{gs} between C_{GO} and C_{GS} given by (651.13) and (651.14) respectively and by using (643.4) and (643.7) for g_d and g_m. This gives

$$C_{gd} = C_oLZ \frac{[3(V_G-V_{GT})-V_D][2(V_G-V_{GT}-V_D)]}{\{3[2(V_G-V_{GT})-V_D]^2\}} \qquad (651.16)$$

which can be shown to be identical to C_{DS} given by (651.9): $C_{gd}=C_{DS}$. Then (651.15) requires that

$$C_{ds} = C_{DS} - C_{gd} = 0.$$

Thus, there is no capacitive coupling between the drain and the source through the channel in the ideal intrinsic one-dimensional MOSFET.

Using C_{gd} from (651.16) and C_{GS} from (651.8) in (651.14) we get

$$C_{gs} = C_{GS} - C_{gd} = C_oLZ \frac{2}{3} \frac{(V_G-V_{GT})[3(V_G-V_{GT})-2V_D]}{[2(V_G-V_{GT})-V_D]^2}. \quad (651.17)$$

The d.c. voltage dependences of the gate and drain capacitances, C_{GS}, C_{GO}, C_{DS}, C_{DO}, and the small-signal equivalent-circuit capacitances, C_{gd} and C_{gs}, are graphed as a function of V_D or V_G-V_{GT} with the other as the constant parameter. The families of curves are given in Figs.651.3 to 651.8.

Asymptotic Results at Low Drain Voltage and Current Saturation

The following simple asymptotic results are obtained at small drain voltages and in the drain current saturation range. They may be used to check the results given in Figs.651.3 to 651.8.

(1) Small drain voltage and below drain saturation voltage. $V_D \ll V_G - V_{GT}$

$C_{GS} = C_oLZ$ and $C_{GO} = C_oLZ$ (651.18)

$C_{DS} = C_oLZ/2$ and $C_{DO} = C_oLZ/4$ (651.19)

$C_{gs} = C_oLZ/2$ (651.20)

$C_{gd} = C_oLZ/2$ (651.21)

(2) Drain current saturation range. $V_D \geq V_G - V_{GT}$

$C_{GS} = C_oLZ(2/3)$ and $C_{GO} = 2C_oLZ$ (651.22)

$C_{DS} = 0$ and $C_{DO} = 0$ (651.23)

$C_{gs} = C_oLZ(2/3)$ (651.24)

$C_{gd} = 0$ (651.25)

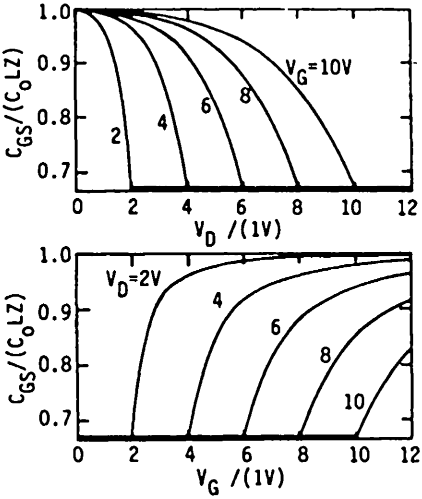


Fig.651.3 C_{GS} vs V_D and V_G ($V_{GT}=0$).

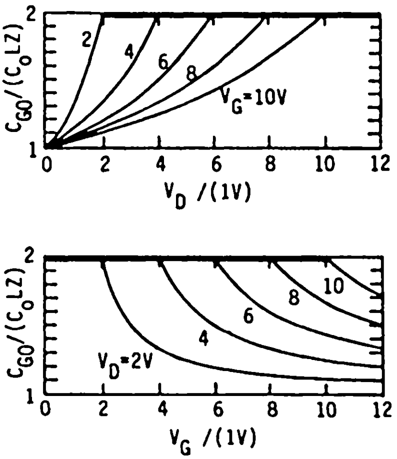


Fig.651.4 C_{GO} vs V_D and V_G ($V_{GT}=0$).

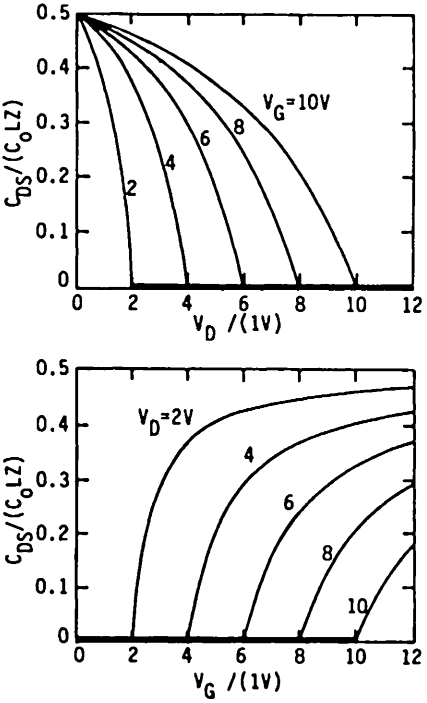


Fig.651.5 C_{DS} vs V_D and V_G ($V_{GT}=0$). $C_{DS}\equiv C_{gd}$.

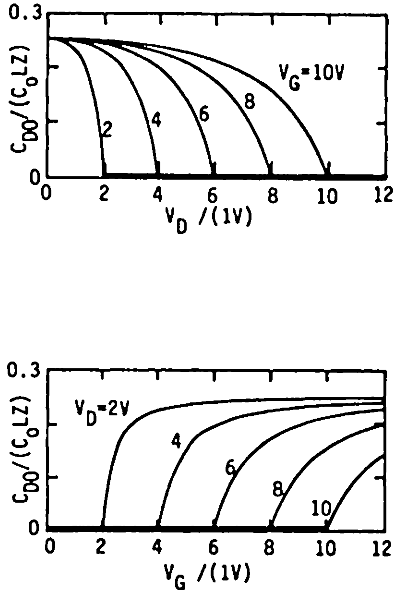


Fig.651.6 C_{DO} vs V_D and V_G ($V_{GT}=0$).

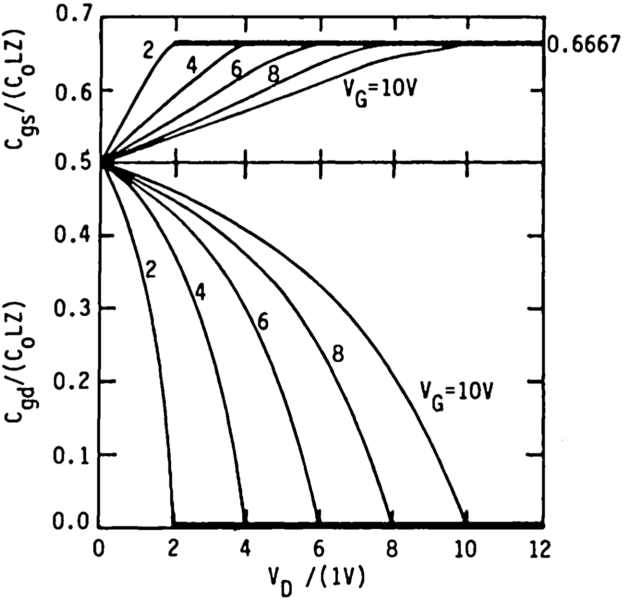


Fig.651.7 Normalized C_{gd} and C_{gs} vs V_D. Note: C_{gd}(here)$\equiv C_{DS}$(in Fig.651.5)

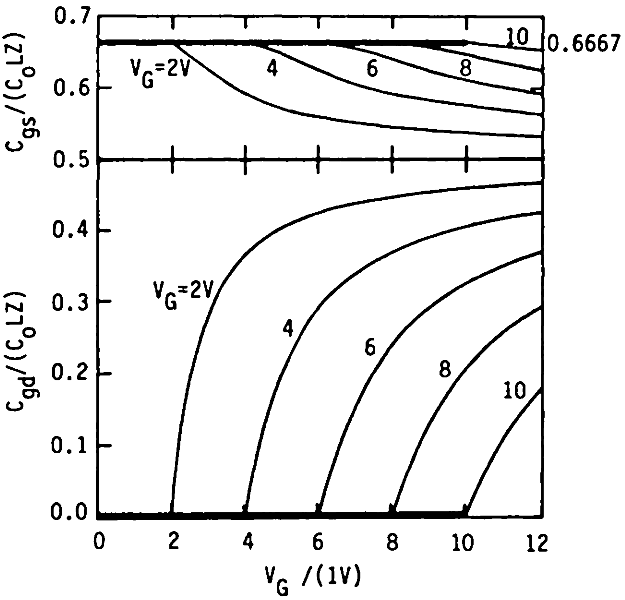


Fig.651.8 Normalized C_{gd} and C_{gs} vs V_G. Note: C_{gd}(here)$\equiv C_{DS}$(in Fig.651.5).

Note that in the saturation range, the feed-through capacitance or the gate to drain capacitance, C_{gd}, is zero in this ideal one-dimensional intrinsic MOSFET. The physical reason is that the carrier charge density at the drain end of the channel (or the $y=L$ position) is zero, $Q_N(y=L)=0$, when $V_D=V_G-V_{GT}$ or when $V_D>V_G-V_{GT}$. This provides the electrostatic isolation between the drain and the gate. The condition of null or depleted carrier density was shown in (642.2) which gave $Q_N(y=L)=C_o[(V_G-V_{GT}) - V(y=L)]=0$ at saturation since $V(y=L)=V_D=V_G-V_{GT}$. This is a very important result which shows that in the simple one-dimensional model, there is no capacitive feedback from the drain to the gate when the drain current is saturated. Thus, any capacitive feedback in a real device would come from two-dimensional effects as well as overlap and parasitic capacitances. The condition of $Q_N(y=L)=0$ is also known as the **pinch off** condition in the theory of field-effect transistors (all types of FET including MOSFET and JGFET as well as others). The descriptive 'pinch off' comes from the physical narrowing of the thickness of the conduction channel in the JGFET device. In the inversion type MOS transistor which we just analyzed, the physical channel is such a thin layer at the oxide-semiconductor interface that its thickness is ill-defined. The word pinch-off must be given a more general meaning in MOST than used in JGET. The broader meaning depicts the pinching off or reducing to zero the conduction charge or conductivity at the drain end of the channel.

652 High-Frequency Response of MOS Transistor

The small-signal high-frequency and high-speed switching responses of MOST's can be estimated from the physics of the mechanisms which delay the output from responding instantaneously to an abrupt change of the input voltage. Two methods will be used: the transconductance cutoff frequency or 3db down frequency, and the gain-bandwidth product.

Transconductance Cutoff Frequency

The transconductance, g_m, has a cutoff frequency since a signal applied to the gate cannot produce an output instantaneously. The delay arises because the drain current response to a signal voltage applied to the gate is carried by the electrons in the channel and it takes a finite amount of time for the electrons to reach the drain from the source. To take into account this small-signal delay in the 1-lump model or charge-control analysis employed here, a single pole approximation for the transconductance may be made:

$$g_m = g_{m0}/(1 + j\omega t_{gm}) = g_{m0}/[1 + j(\omega/\omega_{gm})] \tag{652.1}$$

where g_{m0} is the low frequency transconductance given by (643.7),

$$g_{m0} = (\partial I_D/(\partial V_G)\Big|_{V_D} = (Z/L)\mu_n C_o V_D. \tag{652.2}$$

To derive the exact expression of the transconductance cutoff frequency, ω_{gm}, a distributed or transmission line model must be used even in the one-dimensional approximation employed here. However, a good estimate can be obtained using the charge control analysis. This proceeds as follows. The time delay of the drain current response to a change in the gate voltage or gate charge, or the small-signal delay time, is of the order of

$$\begin{aligned} t_{gm} &= (\delta Q_G/\delta I_D)\Big|_{V_D} = (\delta Q_G/\delta V_G)\Big|_{V_D} \Big/ (\delta I_D/\delta V_G)\Big|_{V_D} \\ &= C_{GS}/g_{m0} = (C_{gs}+C_{gd})/g_{m0} \end{aligned} \tag{652.3}$$

where (651.14) is used for C_{GS}. The transconductance cutoff frequency is then

$$\begin{aligned} \omega_{gm} &= 1/t_{gm} = g_{m0}/C_{GS} = g_{m0}/(C_{gs}+C_{gd}) \\ &= (\mu_n V_D/L^2)\,\frac{3[2(V_G-V_{GT}) - V_D]^2}{3[2(V_G-V_{GT}) - V_D]^2 - V_D^2} \end{aligned} \tag{652.4}$$

or

$$f_{gm} = \frac{\omega_{gm}}{2\pi} = \frac{\mu_n V_D}{2\pi L^2}\cdot\frac{3[2(V_G-V_{GT}) - V_D]^2}{3[2(V_G-V_{GT}) - V_D]^2 - V_D^2}\,. \tag{652.4A}$$

In the drain current saturation range, $V_D \geq V_G - V_{GT}$, then

$$\omega_{gms} = (3/2)(\mu_n/L^2)(V_G-V_{GT}) = 1/t_{gms} \tag{652.5A}$$

and

$$f_{gms} = \omega_{gms}/2\pi = (3/4\pi)(\mu_n/L^2)(V_G-V_{GT}). \tag{652.5B}$$

The normalized small-signal transconductance delay time, $t_{gm}(\mu_n/L^2)$, is plotted as a function of V_D or V_G in Fig.652.9.

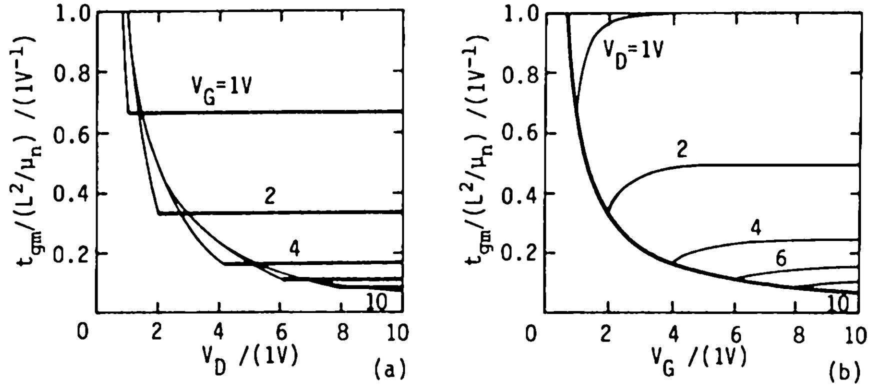


Fig.652.9 Small-signal transconductance delay time of nMOST vs (a) V_D and (b) V_G with V_{GT}=0.

Gain-Bandwidth Product

A second useful figure-of-merit to estimate the high-frequency response of the intrinsic MOSFET, where the parasitic and overlap capacitances are negligible, is the gain-bandwidth product. This quantity will depend on the load impedance that is connected to the output terminals or between the drain and the source. Suppose that the load is purely resistive, given by R_L, and is followed by another MOSFET of identical properties. Then, the voltage gain of one stage is

$$A = -\frac{g_{m0}R_L(1 - j\omega C_{gd}/g_{m0})}{(1 + j\omega/\omega_{gm})(1 + j\omega/\omega_L)} \approx -\frac{g_{m0}R_L}{1 + j\omega/\omega_L} \qquad (652.6)$$

where $\omega_L = 1/R_L(C_{gs}+C_{ds}-C_{gd})$. $R_L = 1/(G+g_d)$ is the effective load resistance. In the intrinsic MOST, $C_{ds}=0$. C_{ds} was carried through the circuit analysis so that (652.6) is valid. If there is a parasitic capacitance loading effect, then C_{ds} is the parasitic capacitance.

The approximate expression in (652.6) is obtained since ω_L is usually the smallest characteristic frequency among the three: ω_L, ω_{gm}, and g_{m0}/C_{gd}. Thus, the - 3db frequency is given by ω_L. The mid-frequency or mid-band gain is given by (652.6) with the signal frequency set at zero. This is

$$A_m = A(\omega\to 0) = -g_{m0}R_L. \qquad (652.7)$$

Using $C_{ds}=0$ for the intrinsic MOST, the gain-bandwidth product is then

$$\begin{aligned}
(\text{gain-bandwidth product}) &= A_m\omega_L/2\pi \\
&= g_{m0}/[2\pi(C_{gs}+C_{ds}-C_{gd})] = g_{m0}/[2\pi(C_{gs}-C_{gd})] \\
&= (3/4\pi)(\mu_n/L^2)[2(V_G-V_{GT}) - V_D].
\end{aligned} \qquad (652.8)$$

In the saturation range, the above product reduces to

$$(\text{gain-bandwidth product}) = (3/4\pi)(\mu_n/L^2)[V_G-V_{GT}]. \qquad (652.9)$$

The general gain-bandwidth product is nearly identical to the transconductance cutoff frequency, f_{gm}, which can be verified by comparing $A_m\omega_L/2\pi$ given in (652.8) with f_{gm} given by (652.4). The difference comes from the feed-through capacitance C_{gd}. This difference disappears at drain current saturation since then the feed-through capacitance is zero as indicated in (651.25).

Because the gain-bandwidth product and the transconductance cutoff frequency are nearly identical, the approximation we made in (652.6) is valid only when the mid- or low-frequency gain is very high or R_L is much larger than $1/g_{m0}$, so that $\omega_L << \omega_{gm}$. The approximation would not be valid if $R_L g_{m0}$ is nearly unity or $R_L \approx 1/g_{m0}$ but this amplifier would not have any gain and is not useful.

653 Numerical Example of Small-Signal Characteristics

From the foregoing charge control analyses and results, we can now compute the numerical values of the conductances and capacitances which are circuit elements of the small-signal equivalent circuit model of the intrinsic MOSFET device. In a realistic transistor, parasitic capacitances and contact resistances must be added onto these intrinsic elements to give a realistic prediction of the small-signal performance of the MOSFET.

Let us consider the device whose d.c. characteristics were computed in section 644. This was an n-channel MOSFET with an oxide thickness of $x_o = 1000A$ and an effective channel conductance mobility for electrons of $\mu_n = 600cm^2/V\text{-sec}$. The oxide capacitance per unit area was

$$C_o = \varepsilon_o/x_o = 3.36\text{x}10^4\ pF/cm^2 = 0.336\ fF/\mu m^2. \tag{653.1}$$

The conductances were given by (643.4) and (643.7) whose numerical values are

$$g_d = (Z/L)\mu_n C_o(V_G - V_{GT} - V_D) = 20(Z/L)(V_G - V_{GT} - V_D)\mu S \tag{653.2}$$

and

$$g_m = (Z/L)\mu_n C_o V_D = 20(Z/L)V_D \quad \mu S. \tag{653.3}$$

The d.c. drain current is given by (643.3) and its numerical value was

$$I_D = (Z/L)\mu_n C_o[(V_G - V_{GT})V_D - ½V_D^2] = 20(Z/L)[(V_G - V_{GT})V_D - ½V_D^2]\mu A. \tag{653.4}$$

For a medium current MOST with a wide and long channel, we used $Z = 20\mu m$, $L = 5\mu m$, and $V_D = V_G - V_{GT} = 5.0V$ at current saturation and had

$$I_D = I_{DS} = 1000\mu A, \quad g_d = 0\ \mu S, \quad g_m = g_{ms} = 400\ \mu S, \quad C_o LZ = 36\ fF.$$

Thus,

$$C_{GS} = C_{gs} = (2/3)\text{x}36fF = 24\ fF, \quad C_{DS} = C_{gd} = 0,$$

and

$$f_{gms} = 2.8\text{x}10^9\ Hz = \text{gain-bandwidth} = 2.8\ \text{gigaHz} = 2.8\ GHz.$$

The very high value of the cutoff frequency or gain-bandwidth product is partly due to the very small gate capacitance of the intrinsic device: C_{GS} is only 24fF or 0.024pF. If we take into account overlap and parasitic capacitances and capacitance loading from the interconnect lines, they can add up easily to 1 pF in a poorly designed device structure with long circuit connections. Then, the figure-of-merit is reduced by 1.0/0.024 = 40 times or to about 70 MHz. This shows the importance and undesirability of the parasitic and overlap capacitances.

654 Distributed Low-Frequency Small-Signal Model

The foregoing analyses of the differential conductances and the charge-control capacitances do not give the resistive components which are in series with the capacitances. An inductance associated with the signal delay in the distributed channel is also missing. These additional circuit elements are shown in the exact first-order hybrid-pi model which has four branches as shown in Fig.654.1. It is a single pole expansion and a good low-frequency approximation which can be used up to the lowest characteristic frequency from the four circuit branches.

The analytical expressions of the circuit elements and characteristic frequencies can be obtained only from the distributed model of the MOSFET via a small-signal expansion of the transmission line equation which gives an integral equation. We shall only list the results and present them in graphs with $V_{GT}=0$. If $V_{GT}\neq 0$, then V_G is replaced by V_G-V_{GT}. The equations and graphs are useful to compute the losses and additional delays in small-signal circuit applications. The equations are listed below. $C_O=C_oLZ=(\varepsilon_o/x_o)LZ$ is the total capacitance.

$$C_{gd} = (2C_O/3)(V_G-V_D)(3V_G-V_D)/(2V_G-V_D)^2 \tag{654.1A}$$

$$g_{gd} = (5\mu_nC_O/2L^2)(V_G-V_D)(2V_G-V_D)(3V_G-V_D)/(5V_G^2-5V_GV_D+V_D^2) \tag{654.1B}$$

$$\omega_{gd} = (3\mu_n/2L^2)(2V_G-V_D)^2/(3V_G-V_D) \tag{654.1C}$$

$$g_{m0} = (\mu_nC_O/L^2)V_D \tag{654.2A}$$

$$\omega_{gm} = (15\mu_n/4L^2)(2V_G-V_D)^3/(5V_G^2-5V_GV_D+V_D^2) \tag{654.2B}$$

$$g_{ds} = (\mu_nC_O/L^2)(V_G-V_D) \tag{654.3A}$$

$$L_{ds} = (4L^4/15\mu_n^2C_O)(5V_G^2-5V_GV_D+V_D^2)/[(V_G-V_D)(2V_G-V_D)^3] \tag{654.3B}$$

$$\omega_{ds} = \omega_{gm} \tag{654.3C}$$

$$C_{gs} = (2C_O/3)V_G(3V_G-2V_D)/(2V_G-V_D)^2 \tag{654.4A}$$

$$g_{gs} = (5\mu_nC_O/2L^2)(V_G(3V_G-2V_D)(2V_G-V_D)/(5V_G^2-5V_GV_D+V_D^2) \tag{654.4B}$$

$$\omega_{gs} = (3\mu_n/2L^2)(2V_G-V_D)^2/(3V_G-2V_D) \tag{654.4C}$$

C_{gd}, g_{m0}, g_{ds}, and C_{gs} are identical to the previous differential and charge-control results. However, ω_{gm} is significantly different. At saturation $V_D=V_G$, the differential result (652.5A) gave $(3/2)(\mu_nV_G/L^2)$, while distributed solution (654.2B) is $(15/4)(\mu_nV_G/L^2)$ or $(15/4)/(3/2)=15/6=2.5$ times larger.

The inductance element in Fig.654.1 is not a real inductance in the usual sense that is associated with a time-varying magnetic field since the very small magnetic field or magnetic force from the conduction current, $v_d\times\mathbf{B}$, is excluded in the Shockley equations. The inductance element is also not a negative capacitance because it would then have a frequency dependence which is not a true circuit element. The origin of the inductance is as follows. It represents the delay at the output or drain node when responding to a signal applied to the gate because the signal carrying electrons must move through the semiconductor RC (dispersive or lossy) transmission line of the surface inversion channel between the source and the drain.

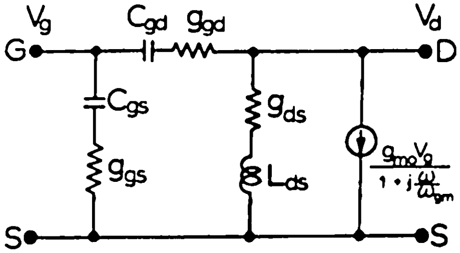


Fig.654.1 The exact first-order small-signal pi model of a MOSFET, substrate and source tied.

The delay is more complex than that of a cascaded multi-section RC filter network because the entire length of the distributed resistive channel is covered by a zero-delay gate conductor over the gate oxide. Thus, the input signal is capacitively coupled simultaneously without delay to both the source and drain ends of the channel (as well as the entire channel). Consequently, the drain output is the sum of the signals of varying delays: the slowest from the source end and the fastest (zero delay) from the drain end. Furthermore, the coupled-in signal at the source end is amplified most because the highest modulation of the channel conductivity by the gate signal occurs at the source end of the channel.

True inductance from the current-induced magnetic field will become important if the current density is extremely high. It can also become important if the geometrical dimensions of the transistor and interconnect circuits are comparable to the wavelength of the high-frequency electromagnetic signal. When the dimension and wavelength are comparable, coupled traveling and standing waves may be set up in the transistor and the interconnect circuits. Then, electromagnetic energy may be radiated by the transistor and the interconnect circuits like an antenna. The dimension and cut-off frequency of present and future VLSI/ULSI/ELSI submicron MOSFETs are approaching the radiation condition. In fact, high-power millimeter-wave monolithic oscillator chips have already been built recently which are composed of 10x10 or larger array of Si MOSFETs and other transistors in a square matrix which are interconnected by radiating transmission lines that serve as the antenna. Electrical tuning of the oscillator frequency is attained by the voltage variation of the drain junction and gate capacitances. Another example of distributed interaction is the junction injection laser which operates under the condition of comparable photon wavelength and junction layer thickness, enabling regenerative feedback that results in coherent emission of light.

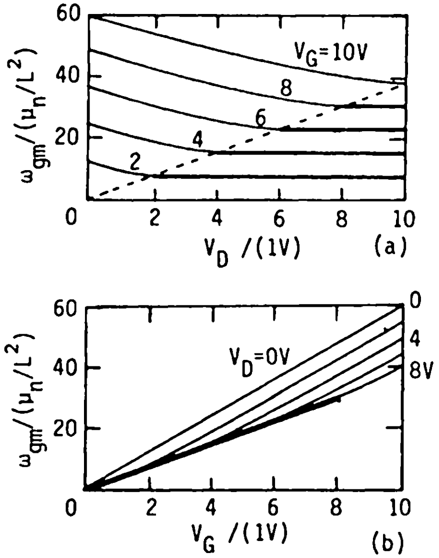


Fig.654.2 The transconductance cutoff frequency ω_{gm} vs (a) V_D and (b) V_G.

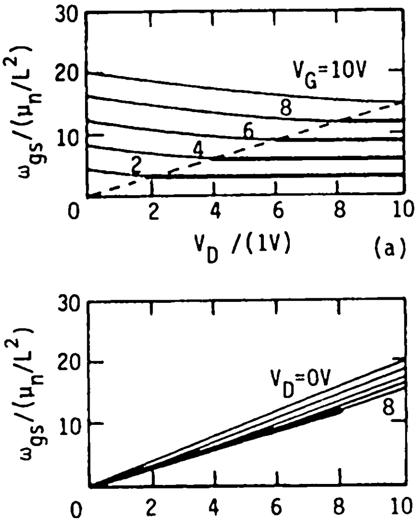


Fig.654.3

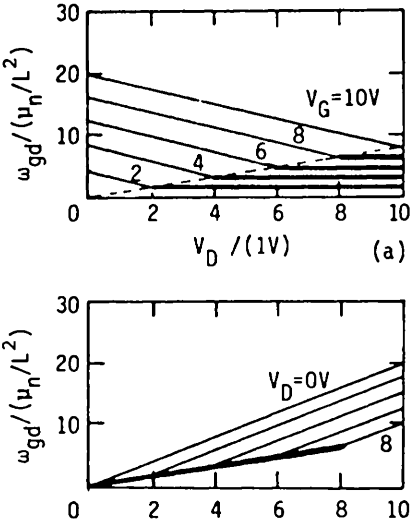


Fig.654.4

Fig.654.3 The gate characteristic frequency ω_{gs} vs (a) V_D and (b) V_G.
Fig.654.4 The feedthrough characteristic frequency ω_{gd} vs (a) V_D and (b) V_G.

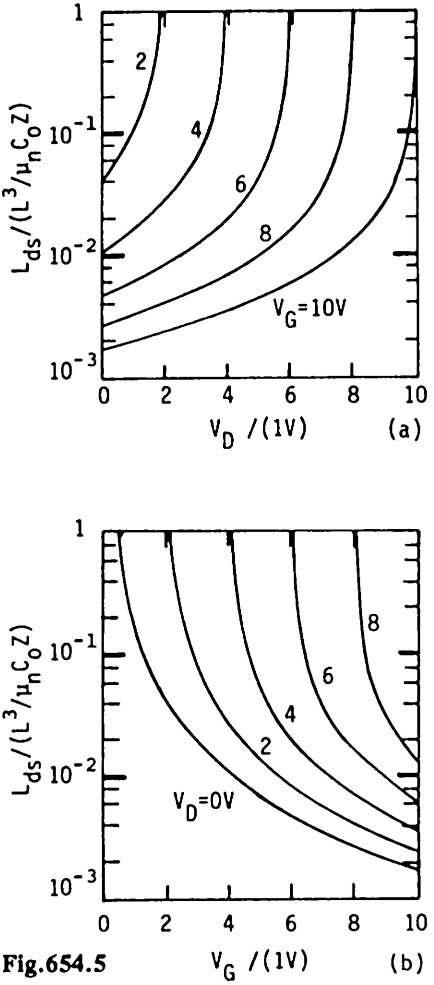


Fig.654.5

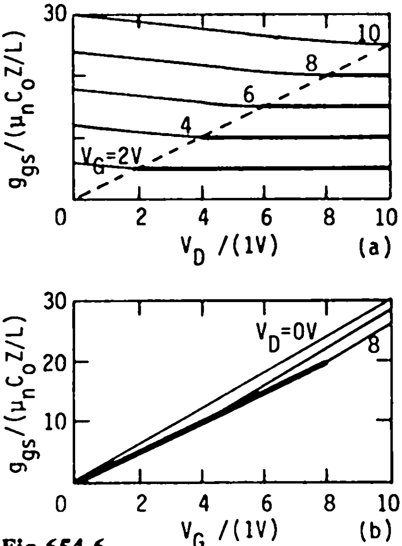


Fig.654.6

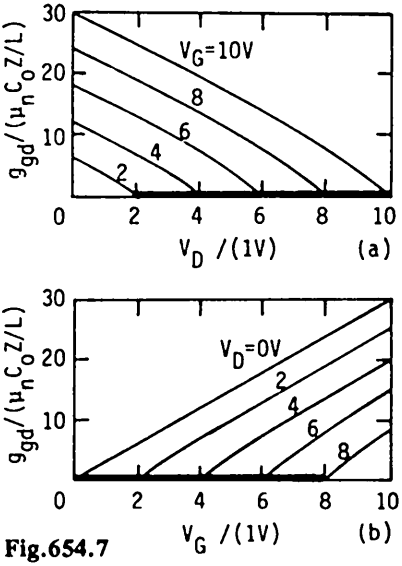


Fig.654.7

Fig.654.5 The channel inductance, L_{ds}, vs (a) V_D and (b) V_G.
Fig.654.6 The gate conductance, g_{gs}, vs (a) V_D and (b) V_G.
Fig.654.7 The feedthrough conductance, g_{gd}, vs (a) V_D and (b) V_G.

660 SWITCHING PROPERTIES OF MOSFET

Silicon MOS transistor is the most abundant object ever manufactured. Its most voluminous applications are in the two types of digital switching circuits, the **memory gate** and the **logic inverter**.

In the **memory gate** and other **gating applications**, the basic building block circuit is just the one MOST (1-transistor or 1T) itself. It is used as the **on-off switch** or **gate** to charge and discharge a capacitor and to transfer charge between two capacitances. The charge gating waveforms will be described in section 663. The DRAM (dynamic random access memory) cells described in sections 671 and 673 store a bit of information in the form of electron (or hole) charges on a charge-storage capacitor or stray capacitance of a node which is gated by a MOST. DRAM cells use four, three or one MOSFETs as the switch or gate to read, write and refresh the stored charges.

In the **logic inverter applications**, the basic building block circuit contains one input-output MOST (1-transistor or 1T) and one load resistor (1R); or two MOSTs (2T), one as the input-output device and the other as the load device. The inverter output voltage waveform on the drain terminal is the **inversion** (or more precisely, the **complement**) of the input voltage waveform on the gate terminal. It is the simplest logic inverter circuit consisting of only **one stage** of the basic block. Complex logic circuits can be built by cascading many stages, including using noninversion stages such as the 1T-1R and 2T emitter follower and the 1T gate. The inverter waveform and switching delay are analyzed in sections 661 and 662. The SRAM (static random access memory) cells consist of two regeneratively cross-coupled inverters, i.e., the output of first inverter is connected to input of the second inverter, and output of the second inverter is connected or fedback to the input of the first inverter. This regenerative feedback or cross-coupling gives bistability or two stable states. It is **latched** onto one of the two stable states until set to the other stable state by a trigger voltage pulse applied to one of the two MOST inverter inputs. It then latches onto the second stable state. SRAM circuits are described in section 673.

The switching speed of MOST digital circuits is determined by two delay mechanisms: the **intrinsic delay** or **transit-time delay** due to transit delay of the electrons (or holes) passing through the channel and the **extrinsic delay** or **load delay** from driving the load and the parasitics by the MOST. These two delays occur simultaneously and are frequently called the **gate delay** which we will infrequently use since we wish to restrict the use of 'gate' to the gating circuit applications described above and to the gate terminal of the MOST.

In monolithic or integrated MOS circuits, the load is mainly capacitive since the gate terminal of a driven MOST in the next stage and the transmission line interconnecting two MOSTs (one driving, the other driven) are both capacitive.

Thus, the **extrinsic delay** comes mainly from two factors: (i) the charging and discharging of the load capacitance driven by the MOST, and (ii) the RC delay of the interconnect lines between the MOST's.

Both intrinsic and extrinsic delays are reduced by shrinking the size of the transistors and integrated interconnect lines in a monolithic chip. But, the resultant increased transistor density (transistor count per unit Si area or per chip) would increase the power density, requiring larger power supply and better heat dissipation. Thus, to evaluate the performance of a MOST design, a second intrinsic figure-of-merit is defined. This is the product of power dissipation and intrinsic delay, known as the **power-delay product**, $P \cdot t_d$, pronounced pee-tee. It is also known as the switched energy, $E_{sw} = Pt_d$. Shrinking the size lowers E_{sw}. But, E_{sw} cannot be smaller than the thermal energy kT in order to limit bit errors arisen from thermal noise owing to random scattering and random generation-recombination-trapping of electrons and holes.

The intrinsic delay, power-delay product, and the extrinsic delays from charging and discharging a capacitance are described and analyzed in the following three subsections, 661, 662 and 663. General analytical expressions are derived. Graphs on the charging and discharging voltage and current waveforms are presented. Numerical examples are also given. An often quoted conclusion is analytically proven: charging up a capacitance by the MOST is slower than discharging - about six times slower, $18(C_L/g_{ms})$ versus $2.945(C_L/g_{ms})$.

661 Intrinsic Delay

The large-signal switching delay of the intrinsic MOST can be estimated in two ways. An intrinsic MOST is the interior portion of the transistor. It excludes overlap capacitances in the device structure. It also excludes the parasitic resistances, capacitances and inductances from the source and drain junctions and contacts, and from the interconnection lines and loads. One of the two methods to estimate the intrinsic delay computes the time required to reach the full channel current after a step gate voltage is applied to turn on the conduction channel. The other computes the transit time of electrons (in n-type channel) from the source to the drain. The two methods give identical large-signal switching time. The large-signal intrinsic delay is expected to be larger than the small-signal intrinsic delay or the reciprocal transconductance cutoff frequency derived in section 652 because the current to charge up the channel is smaller initially when the large gate voltage step is applied. However, the dependence of the large-signal and small-signal intrinsic delays on the device geometry (channel length) and material properties (mobility) are the same as expected.

Consider the switching-time test circuit shown in Fig.661.1. It is in fact the 1T-1R inverter circuit. C_L is the load capacitance. R_L is the load resistance. V_{DD}

is the dc power supply voltage for the drain. The drain current is related to the channel charge by

$$i_D(t) = dq_C(t)/dt. \tag{661.1}$$

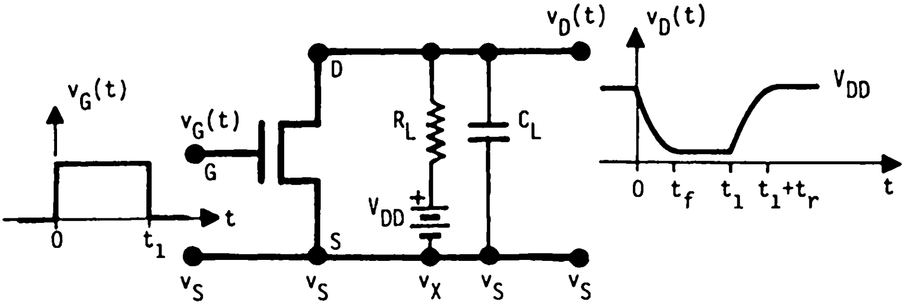


Fig.661.1 The circuit used to calculate and measure the switching delay of a MOS transistor.

A small-signal charging time can be defined by $t_c=\Delta q_C/\Delta i_D$. It is the differential charging time when the drain current is increased by a small gate voltage step. Equation (661.1) also defines a large-signal charging time of the channel which is the time required to charge up the channel from $q_c(t=0)=0$ to the steady-state value Q_C at the steady-state value of the gate voltage, V_G. This large-signal charging time, t_{charge} or t_{ch}, is defined by

$$t_{ch} = \int_0^{t_{ch}} dt = \int_0^{Q_C} dq_C/i_D(t). \tag{661.2}$$

We cannot evaluate the integral exactly since we do not know the exact time dependences of $q_C(t)$ and $i_D(t)$. Thus, we will use the final steady-state value of the current, $i_D(t=\infty)=I_D$, as an approximation. This is a good large-signal approximation since both q_C and i_D are initially zero and both will reach their steady-state value eventually. It could underestimate or overestimate the intrinsic delay depending on the waveform of $i_D(t)$. This is traditionally known as the **quasi-static approximation** since $dq_C(t)$ and $i_D(t)$ at each time interval are approximated by their d.c. static time-independent value. Obviously this is a misnomer according to the accepted usage of the word 'quasi' meaning not-quite true. It should be called **quasi-dynamic approximation** because we are computing a time-dependent dynamic parameter, the delay time, using a static assumption or not-quite dynamic approximation. It is a **piece-wise linear approximation** since $\Delta q_C(t)$ and $i_D(t)$ are assumed to be proportional during the entire switching transient. Then, each proportionality factor (not necessarily constant during the whole transient) or the differential intrinsic delay time, Δt, can be summed to give the total intrinsic delay time. It is also a one-lump-circuit approximation, since the distributed effects does not enter into the derivation of the d.c. charge and current, Q_C and I_D.

Thus, using Q_C from (651.4) and I_D from (653.4) in (661.2) and the quasi-static (quasi-dynamic) piece-wise linear approximation, we have

$$t_{ch} \approx \frac{Q_C}{I_D} = \frac{4L^2}{3\mu_n}\left[\frac{(V_G-V_{GT})^3 - (V_G-V_{GT}-V_D)^3}{[(V_G-V_{GT})^2 - (V_G-V_{GT}-V_D)^2]^2}\right]. \tag{661.3}$$

The second estimate for the large-signal intrinsic delay is the transit time of electrons in the n-channel from the source to the drain. It is given by the differential transit time summed over the length of the channel.

$$t_{tr} = \int dy/v_{drift} = \int dy/[\mu_n|E_y|] = \int dy/[\mu_n(dV/dy)] = \int dy/(I_D/ZQ_N)$$
$$= I_D^{-1}\int_0^L Q_N(y)dyZ \equiv Q_C/I_D \equiv t_{ch}. \tag{661.4}$$

This transit time is exactly equal to the charging time of the channel estimated from the one-lump charge-control analysis, t_{ch}, obtained in (661.2). The equality is not unexpected since the quasi-static (quasi-dynamic) and piece-wise linear approximations are also used.

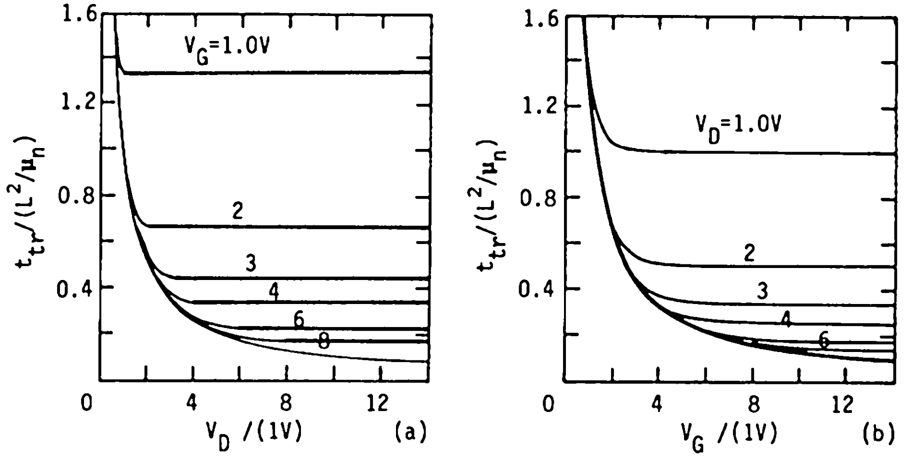


Fig.661.2 The normalized large-signal intrinsic channel charging time constant of MOSTs, $t_{ch}/(L^2/\mu_n)$ ($t_{ch} \equiv t_{tr}$=channel transit time constant). (a) Versus V_D with V_G=constant. (b) Versus V_G with V_D=constant.

The normalized curves of t_{tr} or t_{ch} as a function of V_D and V_G are given in Figs.661.2(a) and (b). Their V_D and V_G dependences are very similar to those of the small-signal t_{gm} shown in Figs.652.9(a) and (b). But the large-signal time constant, t_{tr} or t_{ch}, is larger than the small-signal time constant, t_{gm}, as expected.

At drain current saturation, the large-signal channel charging time or transit time is obtained by letting $V_D=V_G-V_{GT}$ in (661.3), giving

$$t_{chs} = t_{trs} = (4/3)[L^2/\mu_n(V_G-V_{GT})]. \tag{661.5}$$

This is just twice the small-signal intrinsic delay time of the drain current, t_{gms} (or $1/\omega_{gms}$) given by (652.5A) and shown in Figs.652.9(a) and (b). It is expected since charging up the channel from 0 to Q_C (large-signal) by switching $v_G(t)$ from 0 to V_G takes longer than from Q_C to Q_C+dQ_C (small-signal) by switching $v_G(t)$ from V_G to V_G+dV_G, because the charging or transit time is larger at lower channel current or lower gate voltage during the $v_G(t)=0$ to V_G transient.

Using the same mathematical method which gives the analytical solutions of the first order distributed small-signal circuit elements in section 654, the large-signal distributed result in the current saturation range can be derived and is

$$i_D(t) = I_{DS}\tanh^2(\omega_{gms}t/2) \tag{661.6}$$

where $\omega_{gms}=15\mu_n(V_G-V_{GT})/4L^2$ from (654.2B). Thus, at $t=t_{chs}=(661.5)$,

$$\omega_{gms}t_{chs}/2 = (15/4)(4/3)(1/2) = 5/2 = 2.5$$

and

$$i_D(t=t_{chs}) = I_{DS}\tanh^2(2.5) = 0.97340.$$

Delay time is usually defined as the time required to reach 90% of the total response. In this case, it is defined by $i_D(t_{0.9})=0.9I_{DS}$, which gives

$$t_{0.9} = 0.96984[L^2/\mu_n(V_G-V_{GT})].$$

This is 1.3334/0.96984=1.375 times smaller than the transit-time from the charge-control analysis given by (661.5) which is $t_{chs} = t_{trs} = (4/3)[L^2/\mu_n(V_G-V_{GT})] = \underline{1.3334}[L^2/\mu_n(V_G-V_{GT})]$.

The key result is that the intrinsic delay time is approximately given by the channel length divided by the drift velocity. At saturation, $V_D=V_G-V_T$.

Intrinsic Delay Time ~ L/v

$$= L/[\mu_n E_y] = L/[\mu_n(V_D/L)] = L^2/[\mu_n(V_G-V_{GT})] \tag{661.7}$$

Consider the numerical example given in section 644 (x_o=1000A, L=5μm, and μ_n=600 cm^2/V-sec). At V_G-V_{GT}=V_D=5.0V, the intrinsic delay time is

$$t_d \approx t_{chs} = t_{trs} = (4/3)(5\times10^{-4})^2/(600\times5) = 111 \text{ ps.} \qquad (661.8)$$

This intrinsic delay is extremely short. Thus, the switching speed of a digital circuit containing MOST inverters is more likely to be limited by the parasitic capacitances and load resistances of the interconnection lines if the MOST channel is shorter than about 5 μm. For long channel MOSTs, the intrinsic delay or transit time could become comparable to and even larger than the load delay from charging-discharging of the load capacitances. In such cases, the circuit delay time would be controlled by the intrinsic delay of the transistor.

662 Power-Delay Product (Figure of Merit)

The (power-dissipation)$\bullet$(intrinsic-delay) or the **power-delay product** is commonly known as the **figure-of-merit**. This important quantity can be calculated from $\int i_D(t)v_D(t)dt \approx I_DV_Dt_d$ where t_d=t_{tr}=t_{ch}. Using (661.3) for t_{ch}, it is given by

$$\begin{aligned} E_{sv} &= I_DV_Dt_d = V_DQ_C \\ &= C_oLZV_D\left[+ (V_G-V_{GT}) - \frac{3(V_G-V_{GT})V_D - 2V_D^2}{3[2(V_G-V_{GT}) - V_D]}\right]. \end{aligned} \qquad (662.1)$$

In the current saturation range, V_D=V_G-V_{GT}, then

$$E_{sv}(\text{sat}) = (2/3)C_oLZ(V_G-V_{GT})^2. \qquad (662.2)$$

For the numerical example considered in section 644, we have Z=20μm, L=5μm, x_o=1000A, C_oLZ=36fF, μ_n=600cm^2/V-s, V_G-V_{GT}=V_D=5.0V and I_{DS}=1mA. The intrinsic delay computed in (661.8) was 111ps. The power-delay product or the switching energy in the saturation range is then

$$E_{sv}(\text{sat}) = 6\times10^{-13} \text{ Joule} = 600\text{fJ} \qquad (662.3)$$

where 1 fJ or 1 femto Joule is 10^{-15} Joule. This figure-of-merit can be compared with the thermal noise energy, kT, which at T=300K is about 4×10^{-21} Joule or more than one-million times smaller. So, there is still a substantial noise margin.

MOST in current production VLSI chips has a channel length of 1 μm, channel width of 2μm, oxide thickness of 100A, C_oZL=7.2fF, surface electron mobility of 300cm^2/V-s, and operating bias still at 5V. The intrinsic delay is then 111(2/5^2)=9ps and the intrinsic power-delay product is 120fJ. The state-of-the-art (1990) MOSTs reported by the IBM Research Laboratory at Yorktown Heights, New York, has a channel length of about 0.1 μm, a channel width of 0.5 μm, an oxide thickness of about 50A, a mobility of 100cm^2/V-s, and an operating voltage of about 1.0 V. The intrinsic delay is $111(6\text{x}5/50^2)\approx1.33$ps, the oxide capacitance is $C_0ZL=0.345$fF, and the intrinsic power-delay product is $600\text{x}[20/(50\text{x}40\text{x}5^2) =$ 0.24fJ. However, circuit loading capacitances to be discussed in the next section, can increase these values by an order of magnitude or more.

For short channels, the longitudinal electric field in a portion of the channel near the drain junction will be so high to cause drift velocity saturation due to optical phonon scattering discussed in chapter 3. The channel or drain current stops increasing with increasing drain voltage before the drain current reaches the low-field saturation current value, I_{DS} of (643.6), caused by carrier depletion at the point $Q_N(y_d)=0$ or $V(y_d)=V_G$-V_{GT}. In the preceding estimates, drift velocity saturation is neglected. Its inclusion would increase the transit time since v_{drift} in (661.4) is limited to $\theta_{sat}\approx10^7$cm/s. The channel current will be lower for a given drain or gate voltage so that the power dissipation is smaller. The power-delay product will remain about the same due to the opposite changes of the power dissipation and the intrinsic delay.

Three numerical examples on the figure-of-merit are given in Table 662.1. These are the medium geometry MOSFET described in section 654 with a 5μm channel length which is listed on the left column, a current production one-micron MOSFET listed in the middle column, and the 0.1-micron MOSFET. The 0.1-micron channel length is a recent (~1990) theoretical limit estimated by IBM which must be operated at low temperatures (77K liquid nitrogen temperature). Current thought is to immerse the 0.1μm Si MOS CPU chips in a small closed-cycle liquid nitrogen refrigerator in order to realize the higher performance or speed from Si MOSTs. High-performance workstations for 3-d real time graphics and mainframe supercomputers using liquid nitrogen cooled 0.1-micron Si MOS chips could become a reality in the next three (0.5, 0.35 and 0.1μm) or four generations of computer hardware evolution.

Three significant results of the ultimate (0.1μm) MOS transistor may be noted from the rightmost column in Table 662.1: (i) the number of electrons in the inversion channel or on the gate is reduced to 4000 which may require additional circuits and shielding to reduce noise from random access signals, radioactivities due to residual package impurities, and cosmic rays; (ii) the intrinsic transit time, 1.33ps, may be increased by capacitance loading from the interconnect circuits and by propagation delay along the interconnect transmission lines; and (iii) the switched energy, E_{sw}, is reduced to 0.09fJ which is only $2.2\text{x}10^4$ times the intrinsic thermal noise, kT. This signal-to-noise ratio is improved only by a factor of $300/77\approx4$ when the operating temperature is lowered from 300K to 77K.

Table 662.1
Figure of Merit
(Silicon n-Channel MOSTs)

		MEDIUM SIZE Section 644		CURRENT PRODUCTION	FUTURE STATE-OF-ART
Z	=	20	μm,	2	0.2
L	=	5	μm,	1	0.1
x_o	=	1000	A,	100	50
C_oLW	=	33.6	fF,	6.73	0.135
μ_n	=	600	cm^2/V-s,	300	100
V_G	=	5	V	5	1
V_D	=	5	V	5	1
I_{DS}	=	1	mA	2.52	0.0663
g_{ms}	=	400	μS	1004.	134.
Number	=	10^6	electrons	200k	4k *
t_{trs}	=	111	ps	8.88	1.33 **
E_{sw}(sat)	=	555	fJ	112	0.09
kT(300K)	≈	4×10^{-6}fJ		4×10^{-6}	4×10^{-6}
E_{sw}/kT	=	$1.4\times10^{+8}$		$2.8\times10^{+7}$	$2.2\times10^{+4}$

* Too few electrons, noise error.
** Slow down by circuit capacitance loading.

663 Charging and Discharging a Capacitor - Extrinsic Delays

Two unique properties of the MOST have dominated the design of MOS switching circuits. These are: (i) a high input impedance which is capacitive and (ii) a very low leakage current when the MOST channel is turned off. These are exploited to great advantage in MOS circuit designs with the following consequences: the principal function of MOST's in digital circuits is to serve as an on-off switch or gate, and the principal delay is due to charging and discharging the capacitance of a load. The load capacitance consists of the gate capacitance of the following MOST and the transmission line capacitance of the interconnect lines.

Three fundamental capacitance charging and discharging circuits using one MOST gate will be analyzed. These are: (1) the charging circuit, (2) the discharging circuit, and (3) the circuit to transfer charge between two capacitors. They are selected because they are the basic building blocks from which most other capacitance loaded MOST circuits can be derived.

The three basic capacitance charging and discharging circuits are shown in Figs.663.1(a)-(b), Figs.663.2(a)-(b), and Fig.663.2(c). Figures 663.1(a) and (b) give the two identical capacitance charging circuits, commonly known as **pull-up** circuits to circuit designers. Figures 663.2(a) and (b) give the two identical capacitance discharging circuits, commonly known as **pull-down** circuits. The third circuit, shown in Fig.663.2(c), is the general form of the two-capacitance circuit with one capacitor charging up and the other discharging down. It is commonly known as the **charge-transfer** circuit which enables charge transfer from one capacitor to another. It is evident that Fig.663.2(c) reduces to Fig.663.2(a) when $C_D=\infty$ or $C_D>>C_S$ and to Fig.663.2(b) when $C_S=\infty$ or $C_S>>C_D$.

C_S is the capacitance connected between the source terminal and the ground where S stands for source or small. C_D is the drain capacitance connected between the drain terminal and the ground. Frequently, C_L is used instead of C_D where L stands for load or large since C_L may contain the C_D and C_S of many MOSTs plus the interconnect line capacitances. The relative size of C_S and C_D or C_L is determined by the application. For example, the small charge storage capacitance of a DRAM cell has $C_S\approx 36$fF, but it cannot be smaller in order to reduce bit noise error. The load capacitance seen by a MOST in a DRAM and SRAM cell has $C_L\approx 1\rightarrow 10$pF. For example, C_L seen by a MOST **word-line** consists of the gate capacitance of a row of 128 (or more) access MOST gates plus the capacitance of the word line; C_L seen by a MOST **bit-line** driver consists of the drain p/n junction capacitance of a **column** of 128 (or more) access MOST's. In charge transfer (CTD) or charge coupled devices (CCD), such as the CCD camera and the CTD shift register, $C_L=C_S$. They are equal to the sum of the input and output capacitances of the MOST, C_D+C_G, since each MOST is followed by an identical MOST (fan-out of one) and since the diffused source and drain junctions are removed and the removal significantly reduces the capacitances.

The orientation of these circuits in Figs.663.1(a) and (b) and in Figs.663.2(a) and (b) is intentionally rotated 90° from the conventional input/output orientation to show that the two charging circuits are completely identical and the two discharging circuits are also completely identical. This orientation displays the extremely desirable symmetry property of the MOST as a circuit element. The symmetry comes from the arbitrary and hence interchangeable labeling of the source and the drain. Thus, it is not necessary to distinguish a common-source configuration from a common-drain (source-follower) configuration in digital circuits. Even when the source and drain regions are not identical, this circuit symmetry greatly simplifies not only d.c. but also transient analyses. This property has not been noticed by circuit designers, and research and teaching authors.

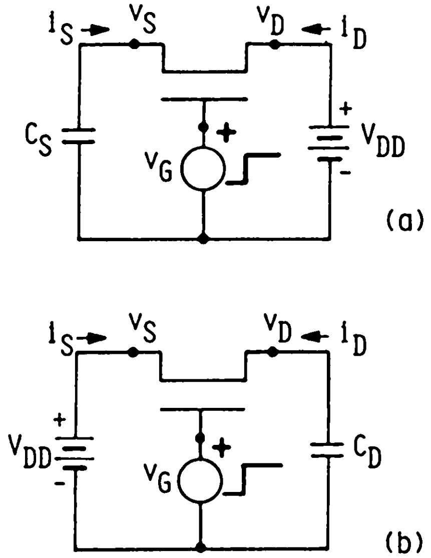

i_S
V_S
V_D
i_D
C_S
V_G
+
V_{DD}
-
(a)
i_S
V_S
V_D
i_D
+
V_{DD}
-
V_G
C_D
(b)

Fig.663.1 The two identical capacitance charging or pull-up circuits by a battery using a MOST gate. (a) Charging a source or small capacitance. (b) Charging a load or large capacitance.

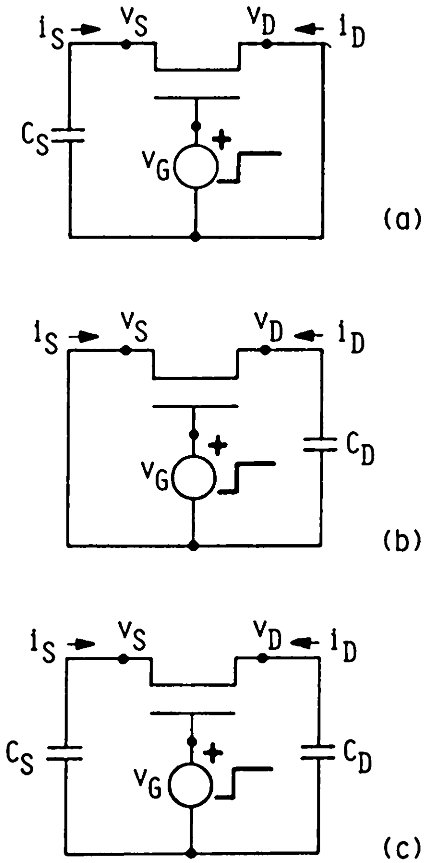

i_S
V_S
V_D
i_D
C_S
V_G
+
(a)
i_S
V_S
V_D
i_D
V_G
+
C_D
(b)
i_S
V_S
V_D
i_D
C_S
V_G
+
C_D
(c)

Fig.663.2 The two identical capacitance discharging or pull-down circuits using a MOST gate. (a) Discharging a source or small capacitance. (b) Discharging a load or large capacitance. (c) Discharging a capacitance onto another capacitance, or charge transfer between two capacitances.

To focus on the circuit delays owing to charging and discharging the interconnect and load capacitors, the intrinsic delay of the MOST discussed in section 661 will be ignored. The MOST is then represented by a resistive or conductive element whose resistance or conductance responds instantaneously to an applied voltage or current transient. This is a particularly good approximation for most digital MOS integrated circuits since they are designed to operate at a lower speed limited by the interconnect and load capacitances instead of the highest speed limited by the intrinsic delay of the MOST. Such a conservative circuit design to operate significantly below the intrinsic speed of the slowest MOST in a dense gate array or memory chip (100k-1M MOSTs) is necessary in order to provide an adequate engineering margin for acceptable manufacturing yields.

When the MOST is designed into a circuit to be switched at the highest possible speed, both the intrinsic and interconnect delays must be taken into account simultaneously in the mathematical analysis. Such a high speed circuit has no exact analytical solution. It can be analyzed only numerically using a circuit analysis and simulation algorithms such as SPICE or its extension to include the RLC delay of the distributed interconnection lines. However, the total circuit delay can be approximated by adding the intrinsic delay to the interconnect delay, but the waveform cannot be predicted accurately. An example is the CTD which in principle can be operated at clock frequencies approaching the intrinsic delay of the drained-sourced MOST's or the drainless-sourceless MOST's since there is no interconnect capacitance loading in a CTD and neither interconnect nor source-drain junction loading in a CCD.

FUNDAMENTAL MOSFET SWITCHING EQUATION

Ignoring the intrinsic or transit time delay given by (661.3) and (661.4) and focusing on the interconnect delay, the MOST can then be represented by a modulated resistance or conductance. Its time-dependent current-voltage equation is exactly the same as its d.c. I-V equation given by (650.1). It can be put into the following symmetrical and time-dependent form. This symmetrical form is especially advantageous for transient analyses of MOS digital circuits as we shall demonstrate with the capacitor charging, discharging and charge transfer examples. We follow the IEEE notation convention which we have discussed and defined in chapter 3 and appendix A. The large-signal time-dependent symmetrical MOST current-voltage equation that possess the drain-source interchangeability is then

$$i_D = -\,i_S = (W/L)(\mu_n C_o/2)(v_{GS}{}^2 - v_{GD}{}^2) \tag{663.1}$$

where

$$v_{GS} = [(v_G - v_{GST}) - v_S]\cdot u(v_G - v_{GST} - v_S) \tag{663.2A}$$

and

$$v_{GD} = [(v_G - v_{GDT}) - v_D]\cdot u(v_G - v_{GDT} - v_D) \tag{663.2B}$$

where u(x) is the unit step function defined by

$$u(x\geq 0) = 1 \qquad u(x<0) = 0. \tag{663.2C}$$

These unit step functions come from the drain (or source) current saturation property. Thus, when the drain voltage exceeds the gate voltage or $v_D \geq v_G\text{-}v_{GDT}$, v_D is set to $v_G\text{-}V_{GST}$ or $v_{GD}=0$. Similarly, when the source voltage exceeds the gate voltage or $v_S \geq v_G\text{-}V_{GST}$, v_S is set to $v_G\text{-}v_{GDT}$ or $v_{GS}=0$.

The currents and voltages are large-signal and time-dependent, for example, $i_D = i_D(t)$ and $v_S = v_S(t)$. The threshold gate voltages are also time-dependent and are measured relative to the source and drain. They include contributions from the fixed oxide traps, chargeable interface traps, the gate-conductor/Si work function difference, and also the substrate bulk charge which comes from the junction-voltage-dependent surface space-charge layer at the source and drain ends of the channel. The substrate charge contribution is derived in a later section and was defined in the MOS capacitance analysis, (411.9). The approximate analytical solutions for the nMOST are

$$v_{GST} = v_{FBS} + 2V_{FS} + \sqrt{2\varepsilon_s q N_{AAS}(2V_{FS}+v_S-v_X)}/C_o \tag{663.2D}$$

$$v_{GDT} = v_{FBD} + 2V_{FD} + \sqrt{2\varepsilon_s q N_{AAD}(2V_{FD}+v_D-v_X)}/C_o \tag{663.2E}$$

For pMOST on n-Si substrate, these equations still apply if appropriate changes of subscripts and sign are made, for example, the sign of the radical due to bulk charge is negative. The extra subscripts S and D in v_{FB}, V_F and N_{AA} are included to account for possible differences of the acceptor dopant concentration and oxide/interface trap densities at the source and drain ends of the channel. For nMOST on p-Si, $v_S(t)\text{-}v_X(t)$ and $v_D(t)\text{-}v_X(t)$ are positive, and v_X is negative if either $v_S=0$ or $v_D=0$, in order to reverse bias the n/p junction of the source, drain and channel.

Additional generalizations can be made to include (i) the junction space-charge layer thickening effect which shortens the length of the channel and (ii) the drift velocity saturation effect near the source and drain junctions which effectively lengthens the channel. These can be included as two ratio multipliers (L/L_{GS}) and (L/L_{GD}) of the two terms in (663.1) to maintain the drain-source symmetry.

$$i_D = -\,i_S = (W/L)(\mu_n C_o/2)[(L/L_{GS})v_{GS}^2 - (L/L_{GD})v_{GD}^2]. \tag{663.3}$$

The two most important general properties of the MOST switch are revealed by (663.1) or (663.3): (i) the drain current saturation and parabolic non-saturation characteristics and (ii) the drain-source symmetry. They not only greatly simplify the analytical mathematical solution of all circuit transient analyses, but also provide a general solution that is the basis of all solutions. These general properties have neither been used systematically to advantage in pedagogical teaching nor in optimizing and systematizing production MOST switching circuit design. These properties will be illustrated with one simple example each, on charging and discharging a capacitance and on transferring charge between two capacitances, using an nMOST as the gate or switch.

Charging a Capacitor

The first unique property just stated concerns the two ranges of the gate-source and gate-drain voltage which was represented by the unit step function in (663.2A)-(663.2C). In the saturation range, the drain or source current is a constant independent of either the drain-gate or source-gate voltage. The MOST is then a **dependent constant current source** which greatly simplifies circuit analyses. For example, consider the case when the channel charge at the source junction is depleted owing to a large voltage applied to the source terminal or a small voltage applied to the gate terminal that makes $v_{SG}(t) \geq 0$ or $v_{GS}(t) \leq 0$ during the entire transient. Then $v_{GS}(t)$ is set to zero and the MOST equation becomes

$$i_D(t) = -(W/L_{GD})(\mu_n C_o/2)v_{GD}^2 = -k_D[v_G - v_{GDT} - v_D(t)]^2. \qquad (663.4)$$

Consider the case of Fig.663.1(b) on charging a load capacitance C_L connected to the drain terminal by a battery, V_{SS}, connected to the source terminal. The charging begins when the nMOST is switched on at $t=0$ by a positive gate voltage step, V_G. For this case of saturated nMOST, V_G must satisfy $v_{GS}(t) \leq 0$ or $v_{GS}(t) = v_G(t) - v_{GDT}(t) - v_S(t) = V_G - v_{GDT}(t) - V_{SS} \leq 0$, giving

$$V_G \leq v_{GDT}(t) + V_{SS}. \qquad (663.5)$$

This is consistent with the practical situation of charging up a bit-line or word-line capacitance of C_L by a MOST in a DRAM or SRAM chip we just described. In practice, V_{SS} may be slightly smaller than V_G such that (663.5) is not valid initially and the MOST is not in the saturation range. But, as the capacitance is charging up, the increasing $v_D(t)$ in $v_{GDT}(t)$ shown in (663.2E) will put the MOST into the saturation range quickly. Thus, the inequality (663.5) is valid except for a short initial duration.

We follow the conventional circuit analysis technique. Thus, the load equation is $i_L(t) = +C_L dv_D(t)/dt$ which gives the circuit equation

$$i_L(t) = C_L dv_D/dt = -i_D(t) = k_D[V_G - V_{GDT} - v_D(t)]^2. \qquad (663.6)$$

The locus of the transient in the $i_D(t)$-$v_D(t)$ plane is shown in Fig.663.3(b). The load or drain current and voltage waveforms, $i_L(t) = -i_D(t)$ and $v_D(t)$, of the simple case to be discussed below are shown in Figs.663.3(c) and (d).

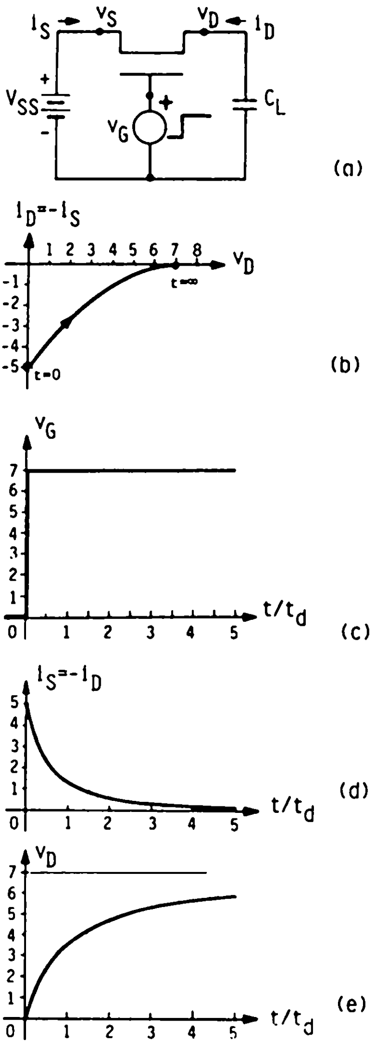

I_S
V_S
V_D
I_D
V_{SS}
+
-
C_L
V_G
+
(a)
$I_D=-I_S$
1 2 3 4 5 6 7 8
V_D
-1
-2
-3
-4
-5
t=0
t=∞
(b)
V_G
7 6 5 4 3 2 1 0
1 2 3 4 5
t/t_d
(c)
$I_S=-I_D$
5 4 3 2 1 0
1 2 3 4 5
t/t_d
(d)
V_D
7 6 5 4 3 2 1 0
1 2 3 4 5
t/t_d
(e)

Fig.663.3

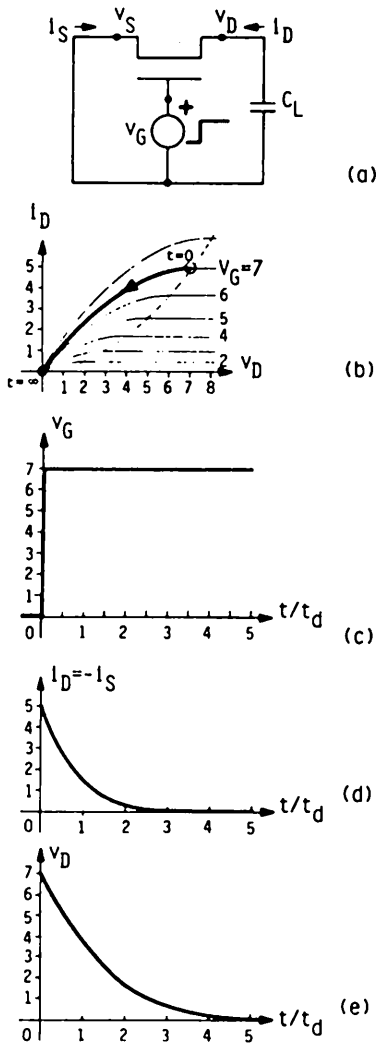

I_S
V_S
V_D
I_D
C_L
V_G
+
(a)
I_D
5 4 3 2 1
t=0
$V_G=7$
6
5
4
2
V_D
t=∞
1 2 3 4 5 6 7 8
(b)
V_G
7 6 5 4 3 2 1 0
1 2 3 4 5
t/t_d
(c)
$I_D=-I_S$
5 4 3 2 1 0
1 2 3 4 5
t/t_d
(d)
V_D
7 6 5 4 3 2 1 0
1 2 3 4 5
t/t_d
(e)

Fig.663.4

Fig.663.3 Charging up a capacitor by an nMOST. **(a)** Circuit diagram. **(b)** Current-voltage locus. **(c)** Applied gate voltage waveform. **(d)** Responding current waveform. **(e)** Responding capacitor voltage waveform.

Fig.663.4 Discharging a capacitor by a nMOST. **(a)** Circuit diagram. **(b)** Current-voltage locus. **(c)** Applied gate voltage waveform. **(d)** Responding current waveform. **(e)** Responding capacitor voltage waveform. Problem: Draw the correct current waveform for (d).

This problem has a very simple solution if we make three time-independence assumptions: (1) the gate threshold voltage relative to the drain terminal is independent of time or $v_D(t)$, $v_{GDT}(t)=V_{GDT}$, (2) the velocity saturation effect is independent of time or $v_D(t)$, and (3) the channel shortening effect is independent of time or $v_D(t)$. The last two makes the effective channel length, L_{GD}, independent of time or $v_D(t)$ and $k_D=K_D=\text{constant}$. These are the very assumptions that have given the simple intrinsic MOST model developed in the preceding analyses and illustrated by numerical examples. The solution of this intrinsic transistor is sublinear in time, i.e., $t/(1+t)$. Linear would be t. The sublinear solution is given by

$$v_D(t) = (V_G-V_{GDT})\,\frac{[(K_D/C_L)(V_G-V_{GDT})]t}{1+[(K_D/C_L)(V_G-V_{GDT})]t} . \qquad (663.7)$$

Note that the capacitance is not charged up to V_{SS} but to V_G-V_{GDT} which is less than V_{SS} or at most equal to V_{SS} according to (663.5). This incomplete charging is due to the assumption of a gate voltage step smaller than the battery voltage, $V_G<V_{SS}+V_{GDT}<V_{SS}$, which puts the MOST in the current saturation range of its i_D-v_D characteristics during the entire transient.

The circuit delay is defined as the time required to complete 90% of the total change. Thus, setting $v_D(t=t_{0.9})=0.9v_D(t\rightarrow\infty)=0.9(V_G - V_{GDT})$ in (663.7),

$$\begin{aligned} t_{0.9} &= 9\{C_L/[K_D(V_G-V_{GT})]\} \\ &= 18\{C_L/[(W/L)(\mu_n C_o)(V_G-V_{GTD})]\} \\ &= 18(C_L/g_{ms}). \end{aligned} \qquad (663.8)$$

Discharging a Capacitor

The other i-v of the first unique property of the MOST is the **non-saturation** range when both v_{GS} and v_{GD} are greater than zero. Physically, no part along the length of the channel is depleted of carrier (or pinched off geometrically if the MOST is a depletion-mode transistor with a chemically doped built-in channel). This is the low-resistance or high-conductance range where the terminal current varies with the terminal voltage **parabolically** in the intrinsic transistor model. It is also known as the triode range named after the non-saturating output current-voltage characteristics of triode vacuum tubes.

As an example, consider the discharging of the bit-line or word-line capacitance, C_L, connected to the drain terminal by a grounded source nMOST shown in Fig.663.2(b) with a high gate voltage step so that V_G is greater than the initial voltage on the capacitor. The condition for nonsaturation during the entire transient is

$$V_G \geq V_{GDT} + V_{DO}. \qquad (663.9)$$

The circuit diagram, current-voltage locus and current-voltage waveforms are shown in Figs.663.4(a)-(e). Substituting $v_S(t)=0$ in (663.1), then, the transistor and load equations are

$$\begin{aligned} i_D &= - i_S = K_S(V_G-V_{GST})^2 - K_D[(V_G-v_{GDT}) - v_D(t)]^2 \\ &= - i_L = C_L dv_D(t)/dt. \qquad (663.10) \end{aligned}$$

Consider the symmetrical case $K_S=K_D=K$ and assume $v_{GDT} = V_{GDT} =$ constant $= V_{GST} = V_{GT}$. The solution can be readily obtained by integration and is given by

$$v_D(t) = V_{DO} \frac{\exp(-t/t_d)}{1 - [V_{DO}/2(V_G-V_{GT})][1 - \exp(-t/t_d)]} . \qquad (663.11)$$

The decay time constant, t_d, is defined by

$$t_d = C_L/g_{ms} = C_L/[(W/L)(\mu_n C_o)(V_G-V_{GT})]. \qquad (663.12)$$

Due to the denominator in (663.11), the voltage decay is super-exponential, i.e., faster than a simple exponential, $\exp(-t/t_d)$.

The 90% response time is the duration required for the capacitor voltage to drop to $0.1V_{DO}$ from V_{DO} by discharge through the MOST. This occurs at the time $t_{0.1}$ given by

$$\begin{aligned} t_{0.1} &= \log_e\{10[2(V_G-V_{GT})-0.1V_{DO}]/[2(V_G-V_{GT})-V_{DO}]\}(C_L/g_{ms}) \qquad (637.13) \\ &= [2.303 + \log_e\{[2(V_G-V_{GT})-0.1V_{DO}]/[2(V_G-V_{GT})-V_{DO}]\}](C_L/g_{ms}) \\ &= [2.303 + 0.642](C_L/g_{ms}) \qquad \text{at } V_G-V_{GT} = V_{DO} \\ &= 2.945(C_L/g_{ms}) \qquad (663.14) \end{aligned}$$

$$\begin{aligned} t_{0.1} &\rightarrow [2.303 + 0 \quad](C_L/g_{ms}) \qquad \text{if } V_G-V_{GT} \gg V_{DO} \\ &= 2.303(C_L/g_{ms}) \qquad (663.15) \end{aligned}$$

Charging-Discharging Comparison

Comparing the capacitance charging time, $t_{0.9}=18(C_L/g_{ms})$, given by (663.8), with the capacitance discharging time, $t_{0.1}=2.945(C_L/g_{ms})$, just obtained in (663.20), it is evident that charging a capacitor by a MOST is much slower than discharging a capacitor by a MOST, in fact $18/2.945=6$ times slower. The reason is revealed by the i-v loci. The charging i-v locus in Fig.663.3(b) shows that the magnitude of the capacitance charging current decreases faster than linear i-v, while the discharging i-v locus in Fig.663.4(b) shows that the capacitance discharging current decreases slower than a linear i-v. The faster-decaying smaller charging current than the slower-decaying larger discharging current makes charging much slower than discharging a capacitor by the MOST.

Charging-Discharging Cycle Time

In some circuit applications, a capacitor is charged and discharged by one MOST or several similar MOSTs. The total charging-discharging cycle, known as the cycle time, is then given by

$$t_{cycle} = (18+2.945)(C_L/g_{ms}) \approx 21(C_L/g_{ms}). \qquad (663.16)$$

For a loading capacitance C_L of 1pF and a MOST transconductance of 100μS, (663.16) gives a 200ns cycle time. This is a typical value for DRAM chips. For a memory storage capacitance of 36fF and an access MOST transconductance of 100μS, the charging-discharging cycle time is 7.2ns if the MOST is not loaded down by line capacitances.

Charge Transferring Between Two Capacitors

The utility of the second unique property of the MOST, its symmetry, can be demonstrated by analyzing the charge-transfer transient between two capacitors, C_S and C_D, connected by a MOST as illustrated in Fig.663.2(c). To illustrate the symmetry, let $C_S=C_D=C_L$, and $V_{GDT}=V_{GST}=V_{GT}$. The circuit, i_D-v_D and i_S-v_S loci, and current and voltage waveforms are shown in Fig.663.5(a)-(e). For convenience, let us set $V_{GT}=0$ which is equivalent to shifting the gate voltage by V_{GT} if $V_{GT}\neq 0$. Then, the device, load and circuit equations are

$$-C_D dv_D(t)/dt = +i_D(t) = C_S dv_S(t)/dt = -i_S(t) \qquad (663.17)$$

$$= (W/L)(\mu_n C_o/2)\cdot\{[V_G-v_S(t)]^2 - [V_G - v_D(t)]^2\} \qquad (663.18A)$$

$$= \tfrac{1}{2}K\cdot\{[V_G-v_S(t)]^2 - [V_G - v_D(t)]^2\} \qquad (663.18B)$$

where $K=(W/L)(\mu_n C_o)$.

The charge is conserved since there is no resistance to dissipate the charge. This charge conservation law can be derived from the equality of the two capacitance currents owing to the two capacitances connected in series. It also gives the relationship between the two unknown voltage variables, $v_S(t)$ and $v_D(t)$, showing that they are complements of each other. From the load equations given by (663.17), we have

$$\begin{aligned} i_D(t) + i_S(t) &= -[C_D dv_D/dt + C_S dv_S/dt] \\ &= -(d/dt)(C_D v_D + C_S v_S) = 0, \qquad (663.19) \end{aligned}$$

which can be integrated to give

$$\begin{aligned} 0 &= C_D[v_D(t) - V_{D0}] + C_S[v_S(t) - V_{S0}] \\ &= C_D v_D(t) + C_S v_S(t) - (C_D V_{D0} + C_S V_{S0}) \end{aligned}$$

or

$$\begin{aligned} Q_0 &= C_D V_{D0} + C_S V_{S0} = C_D v_D + C_S v_S \\ &\equiv \textbf{Charge Conservation.} \qquad (663.20) \end{aligned}$$

The initial capacitance voltages are designated by $v_S(t=0)=V_{S0}$ and $v_D(t=0)=V_{D0}$. Equation (663.20) may be used to eliminate $v_S(t)$ or $v_D(t)$ in (663.18B) to give two differential equations, one for $v_D(t)$ and the other for $v_S(t)$.

Using $C_D=C_S=C_L$ to give complete symmetry in device as well as in circuit, then the charge conservation law gives the **complementary** voltage property

$$v_D(t) + v_S(t) = (V_{D0}+V_{S0}) \equiv V_0. \qquad (663.21)$$

Let us consider the case of $V_G \geq V_0$ so that the MOST is in the non-saturation range during the entire charge transfer transient. Then, from using (663.21) in (663.17) to (663.18B), the two differential equations are given by

$$i_D = -C_L dv_D/dt = \tfrac{1}{2}K[(V_G-v_S)^2-(V_G-v_D)^2] = \tfrac{1}{2}K(2V_G-V_0)(2v_D-V_0), \qquad (663.22)$$

$$i_S = -C_L dv_S/dt = \tfrac{1}{2}K[(V_G-v_D)^2-(V_G-v_S)^2] = \tfrac{1}{2}K(2V_G-V_0)(2v_S-V_0). \qquad (663.23)$$

Note the complete symmetry of these two equations: interchanging the subscripts S and D produces one equation from the other.

These two differential equations show that the transient current is **proportional** to the transient voltage so the loci on the i_D-v_D and i_S-v_S planes are **straight** lines. Figure 663.5(b) shows this straight line locus and the arrow points from $t=0$ to $t\to\infty$.

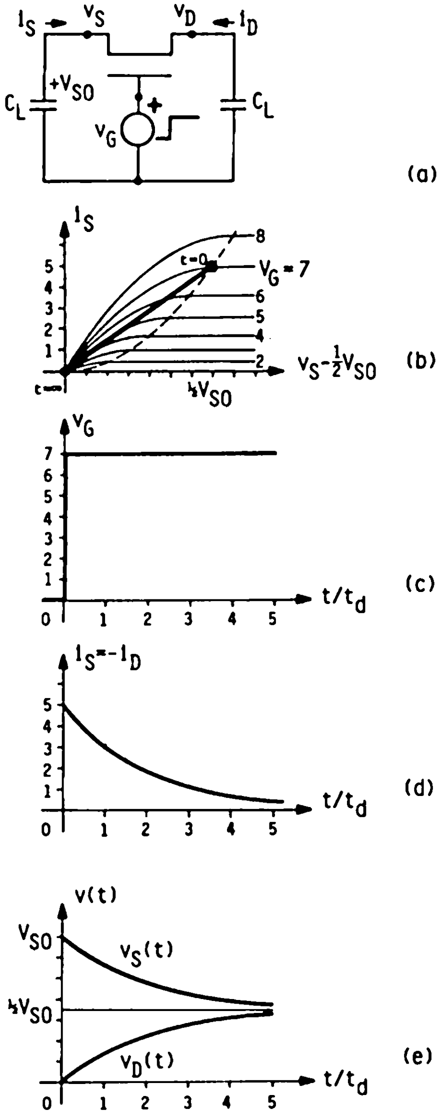

I_S
V_S
V_D
I_D
C_L
$+V_{S0}$
V_G
+
C_L
(a)
I_S
t=0
8
V_G = 7
6
5
4
2
$V_S - \frac{1}{2}V_{S0}$
t=∞
½V_{S0}
(b)
V_G
7
6
5
4
3
2
1
0
1
2
3
4
5
t/t_d
(c)
$I_S = -I_D$
t/t_d
(d)
v(t)
V_{S0}
$v_S(t)$
½V_{S0}
$v_D(t)$
t/t_d
(e)

Fig.663.5 Transfer of charge between two equal capacitors, C_L, gated by an nMOST. (a) Circuit diagram. (b) Current-voltage locus. (c) Applied gate voltage waveform, $V_G = V_{S0}$. (d) Discharging-charging current waveform. (e) Charging and discharging voltage waveforms.

Consider charge transfer from the source capacitance to the drain capacitance. Let the source capacitance have an initial charge C_LV_{S0} or initial voltage of V_{S0} and let the drain capacitance have no initial charge, $C_LV_{D0}=0$ or $V_{D0}=0$. Then the solutions for this case with $V_G \geq V_0 = V_{S0}$ are

$$v_D(t) = \tfrac{1}{2}V_{S0}\ [1 - \exp\{-[1-(V_{S0}/2V_G)](g_{ms}/C_L)t\}] \qquad (663.24)$$

$$v_S(t) = \tfrac{1}{2}V_{S0}\ [1 + \exp\{-[1-(V_{S0}/2V_G)](g_{ms}/C_L)t\}] \qquad (663.25)$$

$$i_S(t) = [1-(V_{S0}/2V_G)](g_{ms}V_{S0})\cdot[\exp\{-[1-(V_{S0}/2V_G)](g_{ms}/C_L)] \qquad (663.26)$$

where $i_S(t) = -\ i_D(t)$ and g_{ms} is the transconductance defined by $g_{ms} = (W/L)(\mu_nC_o)(V_G)$. It is the saturation transconductance. However, the nMOST in this circuit operates in the non-saturation range of the i-v characteristics.

The 90% time constant can be readily evaluated by letting $v_D(t_{0.9}) = 0.9v_D(t=\infty) = 0.9\text{x}(\tfrac{1}{2}V_{S0})$ which gives

$$t_{0.9} = (\log_e 10)[2V_G/(2V_G-V_{S0})](C_L/g_{ms})$$

$$= 2.302585[2V_G/(2V_G-V_{S0})](C_L/g_{ms}) \qquad (663.27)$$

$$= 4.602585(C_L/g_{ms}) \qquad \text{for } V_G=V_{S0}. \qquad (663.28)$$

This is (18/4.605)=3.911 times faster than charging up the capacitance from a battery source, but it is (4.605/2.944)=1.564 times slower than discharging the capacitance C_S to zero. These differences can again be reconciled with the shape of the i-v loci in Figs.663.3(b), 663.4(b) and 663.5(b). A summary is given in Table 663.1.

Table 663.1
Summary of Charging, Discharging and Charge-Transfer Waveforms on a Capacitor Gated by a MOST

TRANSIENT	SPEED	i-v SHAPE		VOLTAGE WAVEFORM
Charging	Slowest	Sublinear (concave parabola)	Fig.663.3(b)	Sublinear T/(1+T)
Transfer	Medium	Linear	Fig.663.5(b)	Exponential ½[1-exp(-½T)]
Discharge	Fastest	Superlinear (convex parabola)	Fig.663.4(b)	Super-exponential exp(-T) + {1-½[1-exp(-T)]} = 2 + [exp(T)+1]

Normalized time: $T=t/t_d$, $t_d=C_L/g_{ms}$

670 CIRCUIT APPLICATIONS OF MOSFET

Several important circuit applications of the MOS transistors will be described in subsections, 67n. The one-function circuits are analyzed since they are the basic building blocks of the multi-function complex monolithic integrated circuits. The descriptions reiterate key features of the material and device physics in order to provide in-depth connections between the fundamental mechanisms of device physics and the transistor characteristics. Introduction of the basic circuits also serves as a bridge between the device physics of this course and the study of more complex multi-function circuits in advanced integrated circuit courses. The choice of the circuit examples is further influenced by the frequency at which they will be encountered in the first engineering design job of an electrical engineering college graduate.

Two preliminaries will be given to facilitate the analyses of these circuits: the transistor circuit symbols and the generalized d.c. I-V equations.

MOST Circuit Symbols - Evolution Chronology

There are many device symbol sets used for the MOST and there is no universal consensus since some sets are more revealing than others in a given circuit diagram, and many symbol sets were designed for specific circuit applications. This author's basic principle for selecting a transistor symbol follows that of the parameter symbols described in chapter 3 (section 398). It is **"all you need to know at one glance"**. It should be the simplest design that reveals at one glance the necessary device physics for a particular circuit application without using a mental or physical translation table.

In general, four circuit symbols are needed in a set in order to represent the four MOSTs: the enhancement- and depletion-mode p- and n-channels. Reduced sets (<4 symbols) with abbreviated symbol diagrams have been used for specific applications. For example, in small-signal applications there may be a structural source/drain asymmetry and the symbols must differentiate the source from the drain. However, in digital circuits, source/drain labeling can be arbitrary to expedite circuit design as demonstrated in sections 660-663. Furthermore, source/drain symmetry in digital circuits is desirable to simplify fabrication processes. Additional simplifications can be made. In an nMOS chip, there is no pMOST so that it is not necessary to use two complex symbols to differentiate an nMOST from a pMOST. If the circuit contains only enhancement MOSTs, then the transistor symbol can be further simplified since it is not necessary to differentiate enhancement and depletion MOSTs. Finally, the substrate terminal can be omitted if the substrate is common to all transistors on the chip or grounded. Frequently the substrate terminal is not shown in a logic or linear circuit diagram even though the substrate terminals of the MOSTs on the chip are isolated by p/n junctions. When omitted, appropriate d.c. biases applied to the substrate terminals are understood. These omissions facilitate the separation of logic or signal design from bias design to ease the design iteration process.

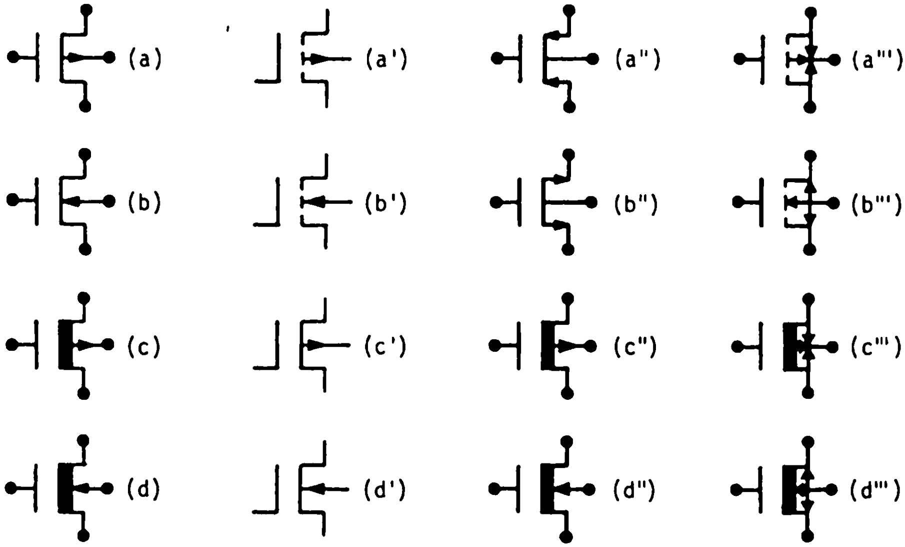


Fig.670.1 Four MOST circuit symbol sets. **(a)-(d)** The simplest (1-p/n diode) four-symbol set. **(a)** Enhancement-mode pMOST. **(b)** Enhancement-mode nMOST. **(c)** Depletion-mode pMOST. **(d)** Depletion-mode nMOST. **(a')-(d')** The corresponding 1979 IEEE standard which is defective (see text). **(a")-(d")** 2-p/n diode extension of (a)-(d). **(a"')-(d"')** The complete and physically correct set showing all three p/n junctions.

Circuit symbols of MOSTs have evolved from the simplest 2-symbol set in 1963 when the CMOS (complementary MOS) inverter circuit was invented, and from the 1-symbol set in 1964 when an Si p-channel enhancement-mode MOST was produced in volume by Fairchild for switching applications and an Si n-channel depletion-mode MOST was produced in volume by RCA for small-signal linear amplification and analog applications [600.1]. The latest symbols used by transistor and circuit designers and research and teaching authors have evolved from them. The simplest (1-arrow or 1-p/n rectifying junction diode) four-symbol set is shown in Fig.670.1(a)-(d). It differs from the set selected by the 1979 IEEE Standards Committee in Figs.670.1(a') to (d') which is defective in three aspects: (i) it is not symmetrical, which obscures the design simplicity owing to source/drain symmetry, (ii) the broken channel line for the enhancement-mode MOSTs takes more effort to draw by hand although it shows the device physics more pictorially and correctly than the unbroken line, and (iii) the thin channel line of the depletion-mode MOSTs in figures (c') and (d') does not give a clear visual impression of an existing inversion or doped channel. Figures (a")-(d") is 2-diode extension of (a)-(d). Figures (a"')-(d"') shows the complete and physically correct 3-diode set which includes all three (drain, source, and channel/substrate) p/n junctions.

A detailed evolutionary chronology of the MOST circuit symbols is shown in Fig.670.2. It contains all the symbols found by the author in a recent literature search. It is by no means complete. In order to facilitate discussion, the circuit symbols are numbered by their coordinate in Fig.670.2, (x,y) or column and row number (c,r). Enhancement, depletion, n- or p-channel is indicated in column -2. The last column, c=x=6, is an extension by this author which gives the two sets of physically most precise but also most complex symbols.

The evolution history of these symbols has not been surveyed by previous authors. Their selection by the originators or inventors was probably not systematic because the fundamental device physics and engineering applications were not visualized all at once in the beginning (1962). The exact historical sequences of evolution of these symbols are probably fairly complex and certainly iterative. Nevertheless, explanations of the symbol evolution can be given based on device physics and intended application, aided by consultation with literature [600.1] and personal recollection.

The symbols in the four rows of the unprimed group, (x=1→5,y=1→4), are perhaps the most compact, revealing, informative, and popular. They evolved from that used by the Fairchild logic circuit designers in 1962, (0,1). The single-primed sets, (x=3→5,y=1'→4'), had evolved from (2,1') which was used by the RCA small-signal linear-circuit designers around 1962. The IEEE 1979 Standards Committee selected the four symbols in the 4th column of the double primed group, (x=4,y=1"→4"), which emphasized asymmetry for linear and analog applications but are undesirable for digital applications. The last group, (x,y=5→8), are symbols used by computer logic circuit designers. They bring out only the necessary features required for expediting the designs of VLSI digital circuit chips. Symbols (0,5), (1,5→8) and (2,5→8), without the substrate lead, are used by IBM engineers. Symbols (3,5→8) are used by DEC (Digital Equipment Corporation) engineers who invented a fifth symbol, (4,6½), for the zero threshold-voltage MOST to facilitate zero-threshold circuit designs.

The very first symbol used in the literature looks like (-2,0). It was designed to show that the conductivity of a semiconductor sheet (the right vertical line) is modulated electrostatically by a gate-conductor plate (the left vertical line). The gate extends beyond the two terminals of the semiconductor in order to effectively modulate the conductance between the two terminals. A modified symbol, (-1,0), was suggested by AMS (American Micro System) engineers in 1970 to show MOST's similarity to a bipolar junction transistor but the slanting of the terminal leads was not adopted. The similarity to bipolar transistor was the cause of initial rejection since the dissimilarity between a majority-carrier field-effect transistor and a minority-carrier-injection bipolar transistor was championed ardroitly by early engineering designers and teachers. But the bipolar-FET similarity turned out to be physically correct when the theory of the MOST at low currents was better understood and derived later. When the MOST is operated at low currents below the threshold voltage, it works like a bipolar transistor because the dominant source-to-drain current is due to the diffusion through the surface channel of the minority carriers injected by the source junction which are collected by the drain junction. The extended gate and substrate lines in (-2,0) were trimmed in the original (1962) Fairchild symbol, (0,1), in order to keep the symbol compact and easy to draw.

It was necessary to distinguish the Fairchild p-channel enhancement-mode MOST from the RCA n-channel depletion-mode MOST when they were marketed in 1964, resulting in the two symbols given by (1,1) and (1,4). The thick line for the base depicts the existing conduction channel at zero gate bias.

The early-day Si n-channel MOSTs were all depletion mode devices because an n-channel exists at zero applied gate voltage. This built-in n-channel was induced by the residual positive oxide charge and the metal/Si work function difference. The zero-bias channel conductivity was further enhanced by boron acceptor outdiffusion into the oxide during oxidation, giving a low boron surface concentration at the SiO_2/Si interface. This built-in n-channel delayed the use of the higher electron mobility than hole to build higher speed n-channel Si MOS logic and memory circuits. It was soon discovered by Heiman at RCA in 1963 that the built-in n-channel on p-Si surface can be eliminated by applying a reverse bias to the semiconductor body (or substrate) relative to the source junction. This discovery, known as the **body effect**, enabled IBM to completely transistorize the mainframe computers in 1973 by replacing the 1-μs magnetic core memory with a 1-μs n-MOS memory. To indicate a d.c. bias applied to the body, a **fourth terminal** was added to the transistor symbol to give the connection to the substrate or body. This is shown in symbols (2,1) and (2,4).

When the CMOS inverter circuit was invented by Wanlass in 1963 (under the direction of research directors V.H.Grinich, G.E.Moore and C.T.Sah at the Fairchild Semiconductor Laboratory), it was necessary to add one more feature to (2,1) in order to distinguish an n-channel from a p-channel. Two 1-arrow designs were proposed by this author. The arrow direction was selected to conform with the IEEE symbol for the rectifying p/n junction diode. This gave the asymmetrical designs shown in (3,1→4). The arrow was located at the drain p/n junction in the 1963 Wanlass-Sah paper which announced the CMOS [600.1] because the source and the substrate are tied which short-circuits the source p/n junction. However, the source p/n junction location for the arrow is better since it indicates both the physical **source** of electrons (or holes) flowing into the channel and also the **channel current direction**. The second 1-arrow design was the symmetrical set shown in (4,1→4) which is physically more accurate because it allows us to apply a (reverse) d.c. bias voltage between the source and the substrate. Fairchild engineers also extensively used the symmetrical design, (4,1→4), as well as (1,1) and (1,4) where the arrow and substrate terminal are dropped to speedup hand- drawing the circuit diagrams since digital MOS circuits need no distinction between the source and drain terminals.

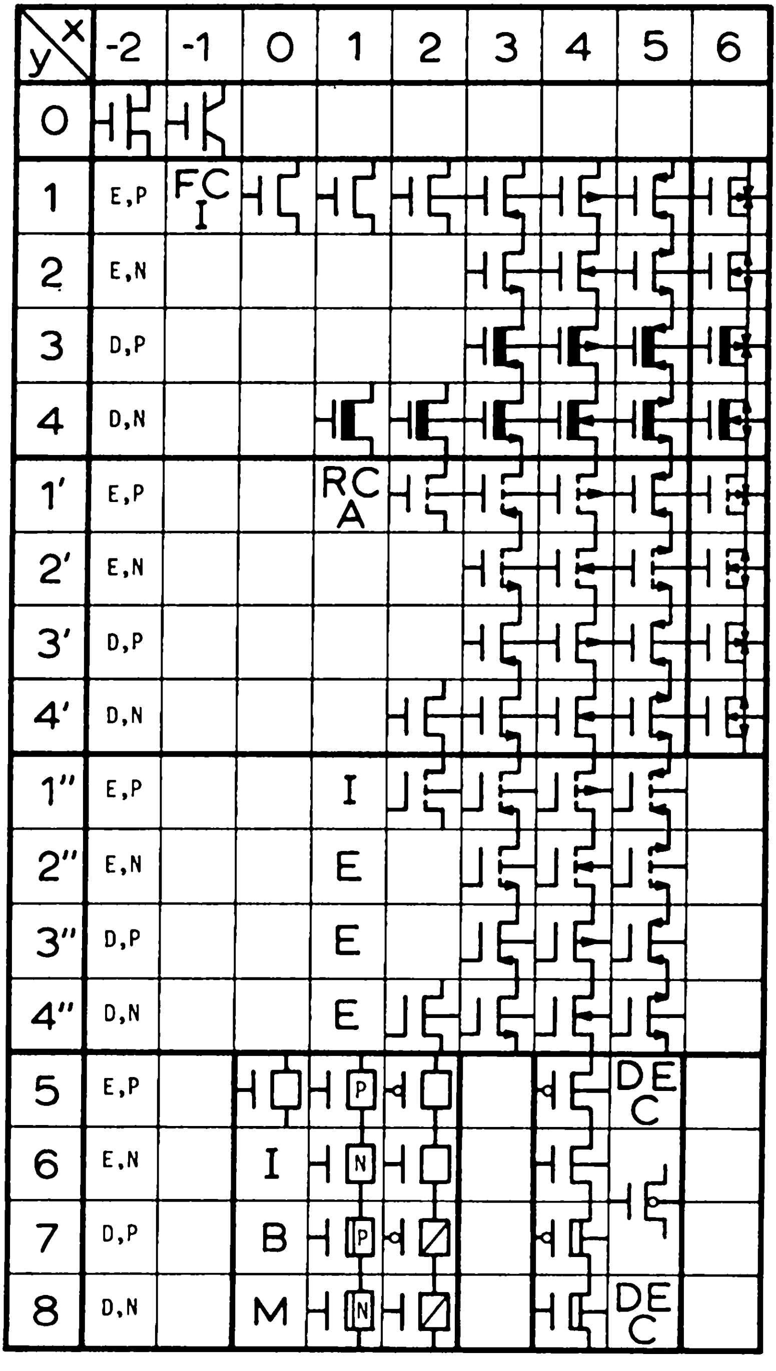

x
y
-2
-1
0
1
2
3
4
5
6
0
1
E,P
FC
I
2
E,N
3
D,P
4
D,N
1′
E,P
RC
A
2′
E,N
3′
D,P
4′
D,N
1″
E,P
I
2″
E,N
E
3″
D,P
E
4″
D,N
E
5
E,P
P
DE
C
6
E,N
I
N
7
D,P
B
P
8
D,N
M
N
DE
C

Fig.670.2 A chronology table of MOST circuit symbols.

An extension of the 1-arrow symmetric MOST symbol is the 2-arrow symbol set shown in (5,1→4) to account for both the source and drain p/n junctions. This is helpful in the d.c. bias design of the BIMOS (Bipolar-MOS) circuit building blocks described in chapter 7. But, it does not show the blocking of the substrate current by the reverse biased surface space-charge layer. This serious drawback is removed by the 3-arrow symbol set shown in (6,1→4).

The RCA symbol sets given by (x=2→5,y=1'-4') and the 1979 IEEE Standard given by (4,1"→4") have two differences from the Fairchild sets just described. One is in the representation of the surface channel. The RCA and IEEE sets use a thin continuous line to represent the built-in channel at zero gate bias for the depletion mode transistors and three disconnected line segments for the absence of a channel in an enhancement mode transistor. The broken line gives the misleading impression that there are three disconnected channels which could be present only in specially designed MOST structures. The spirit of this broken-continuous line symbol set is physically correct, although pictorially somewhat misleading. Another practical hindrance is the difficulty of hand-drawing the short discontinuous line segments. The IEEE 1979 Standard, (4,1"→4"), added another variation by moving the gate terminal lead off-center to the source. This illustrates the asymmetry which is useful only for small-signal common-source circuit applications. Symbol set (3,1"→4") shows another derivative of the RCA set with an off-centered gate lead and the arrow in the source lead which are redundant.

Two other sets of symbols have appeared frequently in the literature because they are used by engineering authors of the two largest computer companies. IBM engineers have used the box symbols shown in (0,5), (1,5→8), and (2,5→8). The channel conductivity type is indicated by the letter P or N inside box, (1,5→8), or by a small circle (depicting holes) for the P-channel, (2,5&7). The built-in channel is represented by a vertical line near the gate side inside the box, (1,7→8), so the letter P or N can still be written conveniently inside the box, or a slanted line inside the box, (2,7→8), where a circle is used to denote P-channel. DEC (Digital Equipment Corporation) engineers have used the symmetrical symbol set shown in (4,5→8). They replaced the arrow in the popular 1-arrow symmetrical set, (4,1→4), by a small circle (for holes) in the gate lead for the p-channel. They also invented a fifth symbol to represent MOSTs with zero threshold voltage by moving the circle to the substrate lead. However, the channel conductivity type becomes ambiguous.

Current-Voltage Equations for MOS Circuit Analysis

In the transient analyses of capacitance charging and discharging by a MOST gate in section 663, it was indicated that the interconnect line capacitance is the limiting factor that slows down the switching speed of MOST digital circuits and not the MOST itself. This was shown to be due to the inherently low transconductance of a field-effect transistor since the circuit delay time is proportional to C_L/g_{ms} where C_L is the load or interconnect line capacitance and

g_{ms} is the transconductance of the MOST. Because the circuit loading capacitance is much larger than (100x) the sum of the intrinsic, overlap and drain n/p junction capacitances of one MOST, $C_L >> C_I = C_{gs} + C_{ol} + C_{ds} + C_{dpn}$, the circuit delay, (C_L/g_m), is much larger than the intrinsic delay, ω_{gm}^{-1} or t_{tr} $(\propto C_I/g_{ms})$. In the MOST circuit examples to be described in the following subsections, 67n, circuit delay due to load capacitance again limits the speed, and the intrinsic delay can be ignored. Thus, the MOST can again be treated as a voltage controlled nonlinear resistance whose d.c. I-V equation can be used to analyze the switching waveforms. This substantially simplifies the mathematics of the transient circuit analysis.

The d.c. and instantaneous equations are listed in (670.2A) to (670.3B). The threshold voltages relative to drain and source are assumed equal and constant in space-time, and denoted by V_T. Four voltage dependent factors are not considered in order to simplify the analytical solutions and to delineate and bring out the device physics. These are: (i) non-constant threshold voltage due to the body voltage and substrate doping, (ii) velocity saturation at high drain or source field, (iii) gate-voltage dependence of surface mobility, and (iv) channel shortening due to the encroaching drain or source junction electric field and the thickening of the junction space-charge layer. These factors were explained but also excluded in section 663 on capacitor charging, discharging and charge-transferring to simplify the analyses.

To reduce the number of symbols in the analyses, a **transconductance parameter**, K, is defined to represent the geometry, mobility and oxide thickness. The definition is

$$K = (W/L)(\mu_s C_o) \tag{670.1}$$

(W/L) is the electrical channel width to length ratio and it is assumed to be equal to the geometrical or photolithographic-mask aspect ratio (Z/Y). μ_s is the average surface mobility of the channel carrier, μ_n for n-channel and μ_p for p-channel. C_o is the gate dielectric capacitance per unit area or oxide capacitance, $C_o = \varepsilon_o / x_o$. All of these are assumed constant to give a constant K. In practice, W, L, and μ_s may vary with the voltages on the four terminals (G,D,S,X). The d.c. MOST equation, (643.2), (650.1) or (663.1), then simplifies to

$$I_D = -I_S$$

$$= (\mu_n C_o Z/2L)\cdot[2(V_G-V_T)(V_D-V_S) - V_D^2+V_S^2] \tag{670.2A}$$

$$= \tfrac{1}{2}K\cdot[(V_G-V_T-V_S)^2 - (V_G-V_T-V_D)^2] \tag{670.2B}$$

$$= \tfrac{1}{2}K\cdot[2(V_G-V_T)-V_D-V_S]\cdot(V_D-V_S) \tag{670.2C}$$

The instantaneous current-voltage equation neglecting intrinsic delay is then

$$i_D(t) = -\ i_S(t)$$

$$= \tfrac{1}{2}K\cdot\{[v_G(t) - V_T - v_S(t)]^2 - [v_G(t) - V_T - v_D(t)]^2\} \qquad (670.3A)$$

$$= \tfrac{1}{2}K\cdot\{2[v_G(t)-V_T] - v_S(t) - v_D(t)\}\cdot[v_D(t) - v_S(t)] \qquad (670.3B)$$

These two **device equations** can then be solved simultaneously with the **circuit equations** from the circuit elements that are connected between the gate, source, and drain terminals. The circuit elements may include resistances, capacitances, inductances, diodes, MOS transistors, and other transistors.

671 Dynamic Random Access Memory Cell, the DRAM

The dynamic random access memory (DRAM) cell, was invented by Robert H. Dennard of IBM in 1967 [600.1]. It contains a capacitor to store the charge (a bit of information) and a MOST switch to access (read and write) the stored charge. It is the simplest integrated-circuit function block and hence most suitable as the first MOS circuit to be studied. It is the easiest high-density (transistor per chip) integrated circuit that can be manufactured at high yields (greater than 90%). Its availability in large volume at low cost (less than 1 milli-cent per bit) have molded the direction of advancement and hastened the application of silicon integrated circuits. It is the most abundant man-made object on this planet earth since its production in volume began in 1972 at Intel Corporation.

Definition of Memory Terms

Concise definitions of memory terms are useful for the discussion of the physics and design of semiconductor memory. These are given as follows.

Memory circuit is an electrical circuit that has **two distinct states** which can be easily distinguished by an external sensor. Then, one **bit** of information can be **stored** in this circuit, and represented by these two states. One state corresponds to the **presence** and the second state, the **absence** of the one bit of information. Thus, the two states can be the two discrete voltage values of a designated circuit node.

Memory cell. The 1-bit memory circuit in a semiconductor memory contains only a few transistors, resistors and capacitors. It is so small, compact and simple that it is known as a cell. The smallest cell is the 1-transistor EPROM cell which is approaching 1 μm^2. Thus, one semiconductor memory cell is a **1-bit store.**

Volatile memory **loses** the stored information when the d.c. power supply is removed. Si DRAM and SRAM are examples.

Nonvolatile memory **retains** the stored information when the d.c. power supply is removed. Magnetic disk and tape, and Si ROM, PROM, EPROM, and EEPROM are examples.

Static memory will work properly and will retain the bit of information during read or access of the stored information without using a clock signal, although clocks are used to synchronize the memory access operations with the other logic operations in a system of many circuits. Si SRAM is an example.

Dynamic memory cannot function properly without a clock signal. The stored information is lost if the clock is interrupted even if the d.c. power is not interrupted. Si DRAM is an example.

The above four distinctions show that there are three classes of memories: the static nonvolatile (magnetic and ferroelectric types, semiconductor capacitor at 77K), static volatile (semiconductor flip-flops), and dynamic volatile (semiconductor capacitor) memories. It is obvious that dynamic and nonvolatile are incompatible.

Read is the operation of sensing the presence or absence of the 1-bit information. In this operation, the state (node voltages) of the memory cell is read.

Destructive Read means that the read operation will change the memory cell to the other state. A **restore** operation must follow a destructive read operation.

Nondestructive Read means that the read operation will not change the memory cell's state.

Write is the operation that sets the state of a memory cell to the '1' state. The 1-state can be represented by a high voltage (+5V) on a two-state node. The 0-state is then the low voltage (0V) on this two-state node. It is evident that the pairing of 1 to +5V and 0 to 0V is not unique, however, it is the most convenient to mentally decipher and understand the operation of a digital circuit compared with the inverse (1 is 0V; 0 is +5V) or the negative logic (1 is -5V; 0 is 0V).

Erase is the operation that sets the state of the memory cell to the '0' state. It is the complement of write.

Program is the write or erase operation of a group of cells, from half-a-byte (4 bits or cells) to 512-1024bits or even all the cells on one memory chip.

Memory chips are manufactured with all or some of the above access capabilities, such as ROM (Read Only Memory), RMM (Read Mostly Memory), WORM (Write Once Read Memory), PROM (Programmable Read Only Memory), EPROM (Erasable Programmable Read Only Memory), UV-EPROM (Ultra Violet Light Erasable Programmable Read Only Memory), EEPROM (Electrical Erasable Programmable Read Only Memory).

Cycle is the combined operation of a read and a write or a read followed by a write. Other necessary operations such as refresh can also be included.

Access time is the time required to read out a randomly chosen bit among the many bits in a memory. It is also used to denote the time required to read out a chosen string of bits from a chip or from a group of chips on a memory printed circuit board. The length of the string can be a byte (1 byte = 8 bits) or a word. A word can be n bytes depending on the computer architecture from the 4-bit (n=1/2) Intel 4004 microprocessor chip to the 64-bit (n=8byte) CDC (Control Data Corporation) and Cray supercomputers. There is no theoretical limit (such as the parallel computers), other than to maximize the speed.

Cycle time is the shortest time a bit or a word can be addressed for read and write without error. It is the sum of the access or read time, the write time, and other time intervals required to retain the information stored in a cell. The cycle time is typically 1 to 2 times the access or read time.

Random access memory (RAM) is a memory whose bits can be accessed randomly. Magnetic disk and core and semiconductor memories are examples.

Sequential memory is a memory whose bits can be accessed quickly only in a restricted sequence. For example, the previous 999 bits must be accessed before one can read, erase, or write the 1000-th bit. Thus, sequential memory is inherently slower than random access memory. Magnetic tape is the most widely used sequential memory. Obviously, random access memories can be programmed and accessed sequentially.

Brief Manufacturing History of DRAM Chip

Journal articles up to and including 1972 were selected by David A. Hodges in a reprint volume, Semiconductor Memories, published by the IEEE Press. The following extends a review given by this author in October 1988 [600.1].

Figure 671.1(a) shows the simplified cross-sectional view of the first DRAM cell used in the 4-kbit ($1k=2^{10}=1024$) Si DRAM chip manufactured by Intel in 1972. The capacitor of this cell structure is laterally separated from the MOST by the diffused source junction. The cell had 8μm linewidth, 5μm effective channel length which is shortened from 8μm by lateral n+ diffusion while forming the n+/p junctions, 1280μm^2 cell area, 1000A gate and cell oxide, and 128fF storage capacitance. The cell area was shrunk in the next two generations (16kb and 64kb chips) by removing the source diffusion, as indicated in Fig.671.1(b) when the **double-poly** technology (double polycrystalline silicon layer deposition technology) was developed to give two overlapping poly-Si conductor layers separated by an insulating poly-Si oxide layer. This was known as the **merged cell**. Another advance was made, in speed, by replacing the large resistance of the long diffused n+ bit-line and high capacitance of the diffused n+/p continuous drain junction of the bit-line with small and isolated n+/p drain junctions. These drain junctions from each cell are connected or **strapped** by an overlaid aluminum **metal bit-line** which has low resistance and low capacitance. n-channel MOST has been used due to the higher electron mobility which gives higher transconductance and lower delay when driving the interconnect line capacitance. The bit density was further increased (to 256kb per chip) by two means: shrinking the linewidth to 2μm and using a higher-dielectric-constant insulator (than SiO_2) for the storage capacitor. Higher dielectric constant is desirable since the cell area is mainly occupied by the storage capacitor and a large storage capacitance is necessary to store sufficient number (10^6) of electrons at 5V to minimize bit error due to circuit noise and radioactivity noise from cosmic rays and traces of residual radioactive elements (Sr^{90}) in the DRAM package material. The first 256k DRAM cell, designed by Hitachi in 1982, is shown in Fig.671.1(c). Two dielectric layers were used: a very thin (about 100A) thermally grown oxide on Si covered by a thin layer (100A) of silicon nitride (Si_3N_4) which not only has a higher dielectric constant (7.5) than the SiO_2 (3.9) but also fills up the pinholes in the very thin SiO_2 layer, thereby increasing the manufacturing yield tremendously. The thermally grown thin SiO_2 layer is necessary to minimize the interface traps at the SiO_2/Si interface by tieing up the silicon dangling bonds. If Si_3N_4 is in direct contact with the Si surface, many Si bonds will be dangling, resulting in high densities of interface traps and causing random, unstable and high threshold voltages as explained in sections 642 and 403. Later-generation 256kb cells utilized composite (CVD over thermal) oxide for storage to eliminate oxide pinholes. The production 1-Mb chips have essentially the same cell design as the 256kb, at smaller linewidth and with thinner oxide. The production 4-Mb chips and the near production 16-Mb chips (1991-summer IBM production date and 1992 production targets by other manufacturers) use a stacked storage capacitor to further reduce cell size and limit chip size. The stack consists of multiple poly-Si layers separated by thin thermal poly-Si oxide layer. A trench etched into the Si is used in the developmental 64-Mb chips (reported at 1991-ISSCC) to limit the chip size. Table 671.1 lists 1-transistor DRAM technology evolution and future trend. (Early history leading to the 1-T DRAM cell is given in Table 673.1 in the SRAM section.) Note, chip area utilization stays at about 50%, half occupied by peripheral circuits due to large area CMOS required to drive the large capacitance of the load and interconnect line to maintain speed. BJT driver will reduce this area. Note also the storage capacitance has decreased to 33pF at 4Mb, 16Mb (5V) and 64Mb (3.3V) and it may drop to 15pF at $\geq$256Mb if bit errors from radioactivity and cosmic rays can be reduced and corrected by on-chip error-correction circuits. The speed is increasing and the access time is decreasing below 30ns due to smaller line width. Below about 0.35μm linewidth, conventional far-field optical lithography may have reached its short wavelength limit unless deep ultra-violet light source and near-field technology are developed. Although electron-beam lithography has been developed that has given <0.1μm transistors, its low throughput (ten 8-inch wafer engraved per hour) compared with optical lithography (60 wafers per hour) makes it uneconomical to produce DRAM chips. Thus, x-ray lithography may have to be used whose development is difficult, tedious, and expensive. Significant delay is expected in the production of 256Mbit and denser DRAM chips if x-ray lithography must be used.

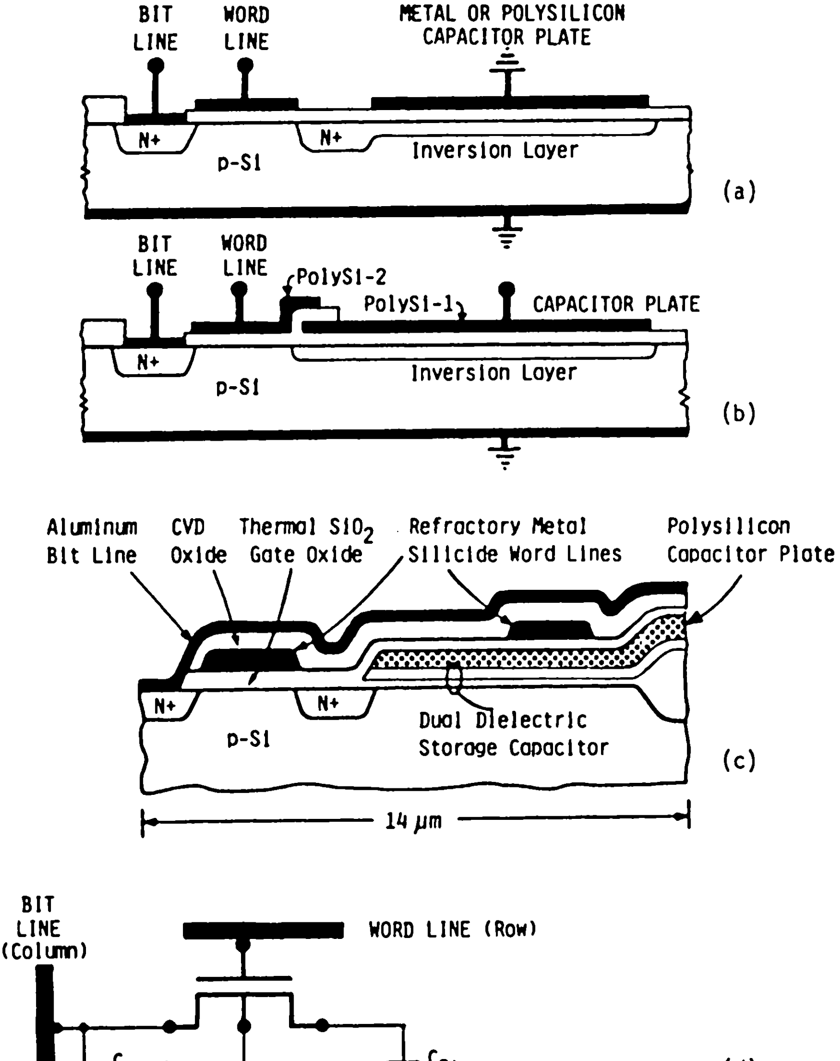

BIT LINE
WORD LINE
METAL OR POLYSILICON CAPACITOR PLATE
N+
p-Si
N+
Inversion Layer
(a)
BIT LINE
WORD LINE
PolySi-2
PolySi-1
CAPACITOR PLATE
N+
p-Si
Inversion Layer
(b)
Aluminum Bit Line
CVD Oxide
Thermal SiO2 Gate Oxide
Refractory Metal Silicide Word Lines
Polysilicon Capacitor Plate
N+
N+
p-Si
Dual Dielectric Storage Capacitor
(c)
14 μm
BIT LINE (Column)
WORD LINE (Row)
C_Load
C_Storage
(d)

Fig.671.1 A sketch of the evolution of the Dennard DRAM cell. (a) The DRAM cell of the first DRAM chip ever produced in volume in 1972 at a cell density of 4k with 8μm feature size. (Intel part number 2104.) (b) The sourceless or merged DRAM cell used in the 16k and 64k DRAM chips (Intel part 2116 for the 16k announced in 1976 with feature size of 5μm. (c) The dual dielectric cell used in the first 256k DRAM announced in 1982). From [600.1]. (d) The circuit model of the DRAM cell.

Table 671.1
Evolution of 1-Transistor Dennard DRAM Cell and Chip

INTR	PEAK	BIT/	LINE	CHANNEL	**OXIDES**		****AREAS*****			SQ/		ACCESS
YEAR	YEAR	CHIP	WIDTH (μm)	LENGTH (μm)	GATE (A)	STORE (A)	CHIP mm^2	CELL μm^2	UTIL (%)	CEL	C_S (fF)	/CYCLE (ns/ns)
****	****	****	*****	*****	****	****	****	****	***	***	****	*******
1972		4k	8	5.	1000	1000		1280			120	
1974	1978	4k					14		36			250/500
1976	1982	16k	5	3.	770	500	17	500	48	60		150/320*
1979	1984	64k	3		500	350	25	180	47		40	150/300
1982		256k	2	1.5	350	200	60	98	42		35	
1985	1987	256k	1.5	1.2	300	200	40	90	56			100/200
1986	1989	1M	1.3	1.0	250	150	55	34	62	30		80/160
1987		4M	1.1	0.9	200	150	99	10.6	43	15	70	70/130
1987	1991	4M	0.8	0.6	200	150	64	7.5	56	9	28	45/ 90
1988		16M	0.7	0.5	~150	~125	148	4.9	53	10	50	80/130
1988	1993	16M	0.7	0.5			94	3.3	56	13	33	65/100
1989		64M	0.5	0.5	100	50	200	1.5	48	~10	33	
1991	1995	64M	0.4	0.5/.7	100	50	176	1.53	50	~10	34	33/R15
		256M	0.25	0.20	100	50	230	0.5	56	7	15?	
	>2000	1G	0.18	0.15	60	40	300	0.15	50	5		

* Load=2 TTL loads + 100pF.

Equivalent Circuit Model of a DRAM Cell

The equivalent circuit of the DRAM cell is shown in Fig.671.1(d). The cell consists of a charge storage capacitor, $C_{Storage}$ (~35fF at 5V to store 10^6 electrons decreasing to 3.3V at 64Mb), a nMOST, and the load capacitance of the bit and word lines, C_{BL} and C_{WL}. The sum $C_{BL}+C_{WL}$ is about 1-10pF. As shown in Fig.671.2(a), it is mostly from the diffused n+/p drain junction of the MOSTs connected to each column of the bit line, and the gate capacitance of the MOSTs connected to each row of the word line.

Cell Array Architecture of a DRAM Chip

The term 'random access' signifies that any cell or bit of information in the chip can be accessed directly without having to access other bits or cells first. The chip architecture gives this property: each cell is located at the cross point of a word line and a bit line, illustrated by the 4x4 cutout in Fig.671.2(a) of the CxR=16x16=256-cells of the DRAM chip in figure (b). Thus, the information stored a group of cells can be accessed in any random sequence when the address (or the x and y coordinates) of each cell is given. The preferred convention is to call the x-coordinate the Column number and the y-coordinate the Row number in order to help memorizing since letter 'C' or x comes before 'R' or y, although some authors and chip designers have interchanged them, causing some confusions.

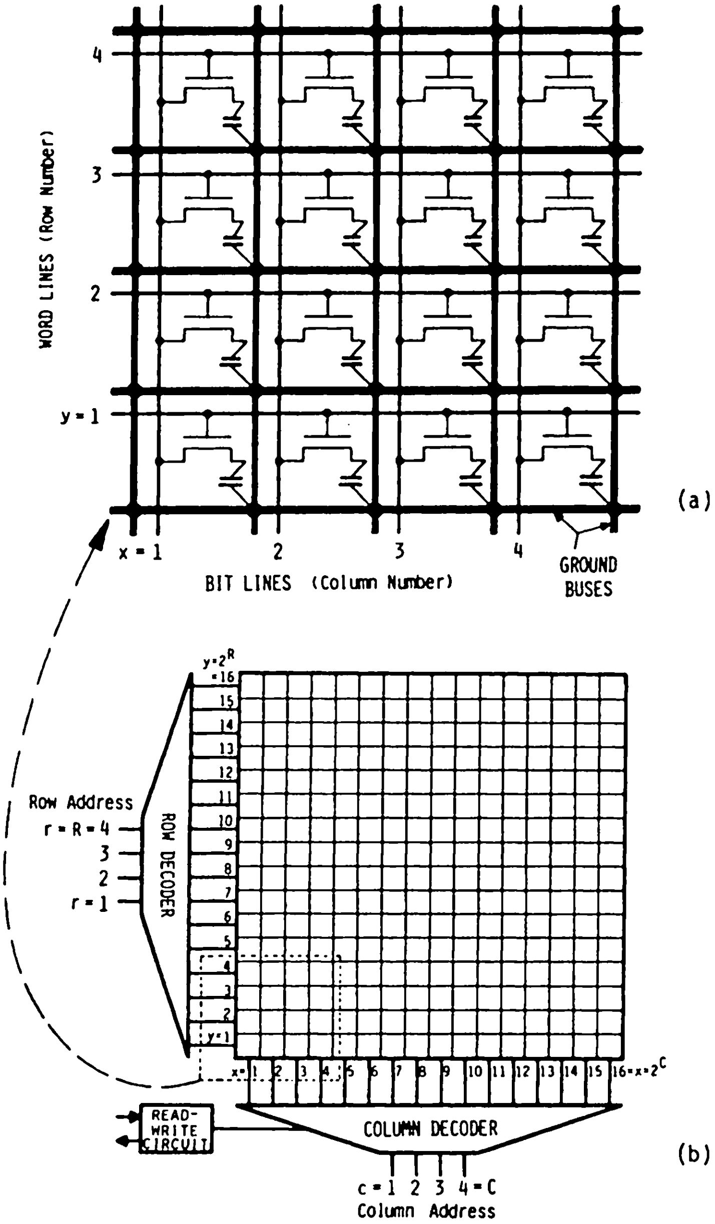

4
3
2
y = 1
WORD LINES (Row Number)
x = 1
2
3
4
BIT LINES (Column Number)
GROUND BUSES
(a)
y=2^R =16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
y=1
Row Address
r = R = 4
3
2
r = 1
ROW DECODER
x=1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16=x=2^C
READ-WRITE CIRCUIT
COLUMN DECODER
c = 1 2 3 4 = C
Column Address
(b)

Fig.671.2 Architecture of a DRAM. (a) A 4x4 array of DRAM cells. (b) Block diagram of a CxR-bit = 16x16 = 256-bit chip with most on-chip circuits omitted.

Column (bit) and row (word) decoders are used to address a bit so the pin counts of the package of a DRAM chip.is a small number, such as the 18- or 22-pin standard dual in-line package. For example, a 1-Mbit chip (2^{20}=1,048,578=1-Mbit) has 2^{10}=1024 column and row lines and would have required 2048 address and data lines without decoders. Figure 671.2(b) shows the block diagrams of a DRAM chip of CxR=16x16=256-bits with x- and y-decoders to reduce lead counts. Other on-chip circuits (not shown) include dummy cells to give balanced signal for lower noise, clock generators, sense amplifiers, read-write drivers, refresh generators, and output buffers. On-chip 5V-to-3.3V voltage divider or 3.3V power supply is needed at shorter channels ($<0.5\mu$m) and thinner gate oxides ($<$ 100A) in order to limit the oxide electric field to less than 1-2 MV/cm to reduce the hot electron effects and to maintain a 10-year operating life. These peripheral circuits occupy about half of the chip area as indicated in Table 671.1.

Basic Operation Principle of DRAM Cell

The basic operation principle of a DRAM cell is described in the following paragraphs. The operation of the many cells of a DRAM chip involves complex timing of the peripheral circuits and is described in advanced circuit books such as Hodges and Jackson (McGraw-Hill, 1988) and Niantsu Wang (Prentice-Hall, 1989).

One bit of information is represented by the presence or absence of electrons on the charge storage capacitor, C_S. Instead of using the electrons in the inversion layer to represent one bit of information, we use the magnitude of the circuit voltage. This allows an immediate connection of the device physics to circuit operation without mental dexterity due to inversion. Thus, we shall use the following convention. **'One' is stored when C_S is charged up to the power supply voltage, $V_{SS}=+5V$.** Charging C_S up to +5V means depleting the electrons in the inversion layer which eliminates the inversion layer. So, the presence of 'one' means the absence of electrons on C_S. The reason of our choice is then obvious: we use the capacitance voltage instead of electron charge so that we do not need to perform the mental inversion of the reciprocal relationships of: '1' (or 10^6 electrons) = 0V, and '0' (or 0 electron) = +5V.

If the charge storage capacitor has a built-in inversion (n-type) layer, then the metal plate of the capacitor can be grounded as shown in Fig.671.1(a). As discussed in chapter 4, such a built-in inversion channel can be obtained by doping the Si surface with a donor impurity, by controlling the metal(or gate-conductor)/Si work function difference, and by having a layer of positively charged oxide traps. Donor doping by ion implantation is preferred since it is reproducible and manufacturable. If a built-in inversion layer is absent, then the metal plate of the charge storage capacitor must be connected to a positive bias voltage relative to the p-Si substrate, such as the +5V power supply, in order to induce an inversion layer. This is shown in Fig.671.1(b) where the capacitor plate is not grounded.

Write 1 Operation

The Write 1 operation involves charging up the C_S to V_{SS} (about +5V) via the bit-line by switching on the MOST with a gate voltage pulse applied through the word-line. The pulse length is made sufficiently long so that C_S is fully charged up to V_{SS}. Thus, the **write delay** involves charging up two large line (word and bit lines) capacitances, discharging the large word line capacitance, and charging up a small charge storage capacitance. The capacitance charging and discharging delays were analyzed in section 663, showing that the charging and discharging delays are respectively $18(C_L/g_{ms})$ and $2.944(C_L/g_{ms}) \approx 3(C_L/g_{ms})$. C_L is the total load capacitance and g_{ms} is the transconductance of the MOST. For $C_L = 1\text{pF}$ and $g_{ms} = 100\mu\text{m}$, $18(C_L/g_{ms}) = 18\text{x}10^{-12}/100\text{x}10^{-6} = 180\text{ns}$. Three design innovations have been developed to minimize the delay and are now described.

(1) Bit Bunching. Cells or bits in DRAM chip can be randomly accessed; every bit or cell can be selected at any time by sending in a pair of addressing pulses on a particular word and bit line. This requires $\sqrt{}$(bit/chip) bit and word lines to access all the cell on a DRAM chip, for example, $2^{10}\text{x}2^{10} = 1024\text{x}1024$ or 1024 bit-lines and 1024 word-lines on a 1-Mbit DRAM chip. This is too many leads. To reduce the lead count, cells are grouped or bunched together in a DRAM chip to limit to 18 or 22 external leads to conform with the industrial standard 18-pin and 22-pin DRAM package. To address each cell or each bit, the write and read signals are multiplexed by on-chip MOST multiplexing circuits. Usually, not more than 64, 128 or 256 MOST's are driven by one word line and one bit line in order to limit the total capacitance of each loaded line to about 1pF. The total load capacitance, C_L, of the loaded line is $C_{\text{world-line}} + 256C_{\text{drain}} \equiv C_{WL}$ for the word line and $C_{\text{bit-line}} + 256C_{\text{gate}} \equiv C_{BL}$ for the bit line in Fig.671.1(d). Bunching into a larger block of bits is known as **partitioning**, for example, a 1-Mbit DRAM can be partitioned into four 256-kbit quadrants on the chip. **(2) Larger n-MOST Line Driver.** The delay can be reduced proportionally by increasing the aspect ratio, W/L, of the driver nMOST since $g_{ms} = (W/L)(\mu_n C_o)(V_G - V_T)$. Current DRAM designs have used CMOS to reduce power dissipation due to the larger n-MOST line driver transistors. Future higher speed DRAM's would employ BiCMOS to further increase the driver transconductance at about the same power dissipation. (3) The bit line is **precharged** to the storage voltage, V_{SS}, to take advantage of the very small delay of charging a very small charge-storage capacitance. For $C_S \approx 32\text{pF}$ (from 10^6 electrons stored at 5V) and $g_{ms} = 100\mu\text{S}$ of the n-MOST gate in each cell, this delay is C_S/g_{ms} - 350ps and negligible compared with the delays of charging up and discharging the 1pF word and bit lines.

Read 1 Operation

The stored charge on C_S (about 35fF and +5V) from the write-1 operation just described is **read** by transferring the stored charge on C_S to the bit-line capacitance C_{BL} and by sensing the charge on the bit line with a MOST sense amplifier. A sense amplifier is needed since the +5V signal on C_S is reduced by the ratio $C_S/(C_{BL}+C_S)$ = 32fF/1pF = 0.032 to 160mV during read. Furthermore, this is a **destructive readout** since the +5V on the C_S is reduced to +160mV after read. Thus, after read, it is necessary to **restore** the voltage on C_S to +5V.

In practical DRAM designs, a differential sense amplifier is used whose reference input is provided by the voltage on a **dummy cell** through another bit line. The differential amplifier reduces the sensitivity to circuit noise from the random currents flowing in the long and closely spaced interconnect lines on the chip. The restore operation is automated during read, and the following sequence of events occurs. Assume that the charge stored on C_S (32pF) gives a voltage of V_{SS} or +5V, and each bit line has C_{BL}=1pF. Then (i) **pre-charge** the dummy cell and the two bit lines (one connected to the storage cell and the other to the dummy cell) to $0.5 \mathrm{x} V_{SS}$ or +2.5V, (ii) turn on the n-MOST via read gate pulse applied to the word line to read, (iii) the voltage on the bit line increases towards the steady-state value of $(C_SV_{SS}+0.5\mathrm{x}C_{BL}V_{SS})/(C_S+C_{BL}) = 0.5\mathrm{x}V_{SS}\{1+[C_S/(C_S+C_{BL})]\}$ = 2.5V+(160/2)mV = 2.5V+80mV with a charge transfer time constant of $3(C_S/g_{ms})$ = 0.96ns since the small C_S is discharging into a very large C_{BL} via the n-MOST, (iv) when the voltage on $C_{BL}+C_S$ approaches 2.5V + 80mV, the very fast cross-coupled differential amplifier is turned on and quickly latches up C_S and C_{BL} to +5V, thereby restoring the C_S from 2.5V+80mV to 5.0V. A similar sequence of events will show that if C_S was initially at 0V, the voltage on $C_{BL}+C_S$ will initially drop towards 2.5V-80mV after read-enable and then will quickly latch to 0V. This is known as **self-restore** or **forced read.**

Further improvement in noise immunity has been achieved by placing the signal and dummy bit lines in parallel on the chip (known as **folded bit-lines**), instead of the original co-linear geometry (known as **open bit-lines**). Common mode noise on the two tightly spaced parallel bit-lines is then rejected by the differential sense amplifier.

Refresh Cycle and Origin of 'Dynamic'

The stored charge on C_S leaks off with time due to recombination and generation of electrons and holes in the Si surface space-charge layer connected to one plate of the charge storage capacitor, C_S. Periodic refresh controlled by a clock must be made in which +5V is written to C_S if it is in the 1-state and 0V is written to C_S if it is in the 0-state. Obviously, the regenerative differential sense amplifier used in the force read operation is precisely the circuit that can perform the refresh operation. Since a string of 256, 128 or 64 cells is tied to one bit-line and one word-line, one would refresh all the cells tied to one word-line by performing a simultaneous read operation to all the cells tied to the selected word line. The refresh time interval is about 200ms to be estimated in the next paragraph. Thus, for a cycle time (read-write) of 200ns, the refresh time interval is about 128x200ns=25.6μs and the refresh operation reduces the DRAM speed by about 200ns/2ms=1%.

The 2 millisecond refresh interval can be estimated from the underlying physics on thermal recombination-generation in the space-charge layer discussed in chapter 3, and the device examples in chapter 4 (MOSC transient), and chapter 5 (p/n junctions). Electrons in the n-type surface inversion channel, which is one plate of the capacitor, will recombine with the holes from the p-type Si bulk. The recombination takes place at the defect and impurity centers or traps in the surface space-charge layer next to the inversion layer. Similarly, if C_S is empty of electrons (+5V) it would not stay empty indefinitely due to the thermal generation of electrons and holes at these traps in the surface space-charge layer. These are the very mechanisms that gives the SNS generation-recombination current in the space-charge layer of p/n junction diodes, J_{SNS} of (535.4), and the capacitance and current transients in MOSC and p/n diodes discussed in sections 421, 422, and 554. Thus, DRAM cells require a periodic refresh to recharge or redischarge the capacitors. The refresh interval is estimated theoretically as follows. The thermal generation rate of electrons and holes at the traps in an Si space-charge layer is given by $e_n^t e_p^t/(e_n^t+e_p^t)\approx e_p^t$ which is about 1000 electrons and holes per second at room temperature if the trap energy level is located near the Si midgap, $E_T\approx E_I$. The trap/dopant concentration ratio, N_{TT}/N_{AA}, can be made as low as 10^{-4} or smaller if the fabrication processing steps are clean and free of thermal shocks to minimize the formation of physical defects (vacancies and divacancies) in the surface space-charge layer. Then, the charging and discharging rate of the inversion layer from the MOSC phase II time constant in section 421 is $(N_{TT}/N_{AA})(1/e_p^t)\approx 10^{-6}$ electron/s. Thus, it is necessary to periodically refresh the stored charge in each cell at about one millisecond interval to limit the fluctuation of the 10^6 stored electrons on C_S to less than 0.1%. The thermal generation rate is thermally activated with a temperature dependence of $\exp(-E_G/2kT)$ from the n_i dependence of J_{SNS}. Thus, the required refresh rate increases rapidly with temperature which increases the cycle time and slows down the DRAM chip.

The number of electrons stored on C_S must be sufficiently large so that it is not disturbed by circuit (current) noise and by randomly generated electrons from residual radioactive elements in the packing material and from cosmic rays. The error from these random noise sources is known as **soft error** since it is a transient and non-permanent error in contrast to the permanent **hard error** due to aging which would permanently disable the chip. About one million electrons are needed to minimize the soft error. This criterion gives the necessary input data to compute the minimum area of the charge storage capacitance and hence the cell area or chip density since $Q=C_SV_S$ and V_S is limited to the power supply voltage or 5 V. Thus, the stored charge is $10^6\text{x}1.6\text{x}10^{-19}=CV=C\text{x}5$, giving a C_S of

$$C_S = 10^6\text{x}1.6\text{x}10^{-19}/5 = 32\text{fF}. \qquad (671.1)$$

Assuming a 200A storage oxide, the storage capacitance area is then

$$\begin{aligned}\mathrm{Area}(C_S) &= C_S x_o/\varepsilon_o = 32\times10^{-15}\times200\times10^{-8}/3.9\times8.85\times10^{-14}\\ &= 18.54\ \mu\mathrm{m}^2 = (4.3\mu\mathrm{m})^2. \qquad (671.2)\end{aligned}$$

Such an area would limit the cell density (cell/chip) to about 1-Mbit or 10^6 cells without unduly large chip since larger chip lowers chip/wafer count and yield, and increases cost. The area occupied by 1M cells is about $(4.3\mathrm{mm})^2 \approx 20\mathrm{mm}^2$. The chip area is twice larger, $\sim 40\mathrm{mm}^2$, to accommodate the peripheral circuits. These are consistent with the production data given in Table 671.1, $34\mu\mathrm{m}^2$ and $55\mathrm{mm}^2$. Thus, to limit the chip size in higher density chips (4Mb, 16Mb, 64Mb and higher), three-dimensional cell designs are needed to reduce cell area. This is accomplished by making the charge storage capacitor a multilayer dielectric-conductor sandwich or composite SiO_2 that is stacked on top of the MOS transistor. Cell designs with trenched capacitor dug into the Si may also be used to reduce cell area.

To limit chip area, small-area bipolar transistors may be used as peripheral drivers to replace the large area inherently low transconductance MOST drivers. Speed increases of 5 to 10 times have been realized. This technology is known as BiMOS DRAM (Bipolar MOS) and BiCMOS DRAM. CMOS is discussed in the last part of the following section, and BiMOS and BiCMOS, in section 766.

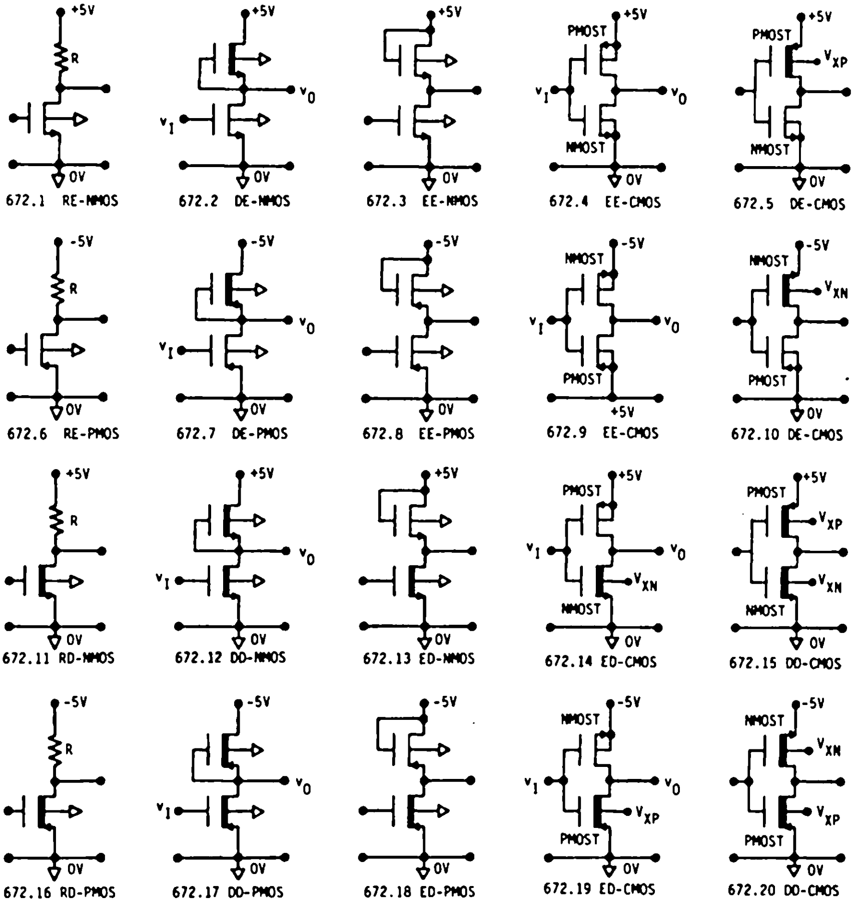
672 The MOS Inverter Circuits

The output waveform of an inverter circuit is the inverse of the input waveform. The **MOS inverter** uses a MOS transistor as the **input device**, known also as the drive device. A resistor or another MOST is used as the **internal load** or the **load device**. The MOS inverter is one of two **most basic building blocks** of all MOS digital integrated circuits. The other is the **charge transfer gate** discussed in section 663 in which the MOST is the switch through which the charge is transferred between two capacitors.

Encyclopedia of Twenty MOS Inverter Circuits

There are twenty possible MOS inverter circuits. They are constructed by selecting one of the four MOSTs (enhancement and depletion nMOST and pMOST) as the input device, and one of the five different loads: the four MOST loads and the semiconductor or thin metal film resistor load. They are shown in Fig.672.1 to Fig.672.20. To help reading the current direction and the d.c. bias polarity at one glance, the rectifier diode arrow in the symmetrical MOST symbol described in Figs.670.1(a)-(d") is moved from the substrate (body) terminal to the source terminal as indicated in Fig.670.2(x=3,y=1→4). The direction of the arrow, still pointing from p-Si to n-Si consistent with the p/n rectifying junction convention, is now also the **direction of positive channel current**. This not only shows, at one

glance, the direction of the channel current but also the bias polarity as well as the high and low voltage states of the input, output and internal nodes. The source location of the arrow also conforms with the emitter location of BJT (chapter 7) and its physics: the n+ source and n+ emitter are the source of the electrons that produce the load current. The drain arrow location used by Wanlass and Sah in their 1962 CMOS invention article will be discarded because it causes confusion: the arrow is opposite to the direction of positive channel current.



672.1 RE-NMOS 672.2 DE-NMOS 672.3 EE-NMOS 672.4 EE-CMOS 672.5 DE-CMOS

672.6 RE-PMOS 672.7 DE-PMOS 672.8 EE-PMOS 672.9 EE-CMOS 672.10 DE-CMOS

672.11 RD-NMOS 672.12 DD-NMOS 672.13 ED-NMOS 672.14 ED-CMOS 672.15 DD-CMOS

672.16 RD-PMOS 672.17 DD-PMOS 672.18 ED-PMOS 672.19 ED-CMOS 672.20 DD-CMOS

Fig.672.1 to Fig.672.20 The twenty possible MOS inverter circuits using one of the four MOSTs as the input device and a resistor or one of the four MOSTs as the load device. Note the CMOS arrow here is at the source rather than the drain p/n junction which was used in the 1963 CMOS invention article of Wanlass and Sah [600.1] and had caused significant confusions.

The five different loads are used to distinguish the inverter circuits, and to coin the acronym. The following are the first five examples.

Fig.672.1	**RE-NMOS**	Resistance-load	Enhancement-input nMOST
Fig.672.2	**DE-NMOS**	Depletion-load nMOST	Enhancement-input nMOST
Fig.672.3	**EE-NMOS**	Enhancement-load nMOST	Enhancement-input nMOST
Fig.672.4	**EE-CMOS**	Enhancement-load pMOST	Enhancement-input nMOST
Fig.672.5	**DE-CMOS**	Depletion-load pMOST	Enhancement-input nMOST

The first three inverters are frequently abbreviated as the **NMOS** circuits and **NMOS Technology** since they employ only nMOSTs. The corresponding three inverters using pMOST shown in Figs.672.6 to 672.8 are known as **PMOS** circuits or **PMOS Technology**. Note that NMOS and PMOS are circuits consisting of one or more transistors, resistors, and capacitors, while NMOST or nMOST and PMOST or pMOST are acronyms for single transistors.

The two remaining inverters (Figs.672.4 and 672.5) utilize a MOST load whose channel conductivity type is the **opposite** to that of the input MOST. They are known as **CMOS (Complementary MOS)**. The two CMOS' are distinguished by their load MOST. The most frequently used CMOS inverter is the EECMOS with the enhancement MOST load shown in Fig.672.4. This CMOS was invented to take advantage of its zero or extremely low standby power dissipation. {See the Wanlass and Sah article presented at the 1963 ISSCC and cited in reference 90 of [600.1].} In this CMOS or EECMOS, both the input and load MOSTs are nonconducting in the quiescent or standby state. Current is drawn only during the switching transient. The other CMOS, the DECMOS shown in Fig.672.5, dissipates power during quiescent or standby state since the D-pMOST conducts. Aside from standby power dissipation, the built-in channel of the depletion pMOST (or depletion nMOST in Fig.672.10) is more difficult to reproduce and requires more processing steps which increase the manufacturing cost. Thus, the acronym CMOS is universally used to denote the EECMOS. In fact, the theory and performance analyses of the DECMOS, EDCMOS and DDCMOS were given only recently (around 1985). Their minor performance advantages over the EECMOS are not sufficient to compensate for their fabrication complexity and higher manufacturing cost as well as large standby power dissipation.

One interesting feature of the CMOS inverter is that one can split the power supply between the top and bottom source terminals, as shown in Fig.672.9, to give a balance or differential output waveform whose two states are +5V and -5V. However, most MOS integrated circuits use only one +5V power supply, with the source of the nMOST grounded and the source of the pMOST connected to +5V. It is also evident from Fig.672.5 that the substrate or body terminal must be shorted to the source to prevent the source junction from being forward biased either by the d.c. power supply or a switching transient. The selection of the MOST symbol with the arrow moved to the source junction facilitates the adherence of this bias circuit design rule.

The most important and widely used MOS inverter circuits utilize the enhancement-mode nMOST (Figs.672.1 to 672.4) or pMOST (Figs.672.6 to 672.9) as the input device so that the input transistor is normally off without having to use a d.c. gate bias to cut off the input transistor. When the input MOST is off or nonconducting, there is no current flowing through the resistor or transistor load, hence, the output is at the full power supply voltage, V_{DD} or V_{SS}, (for the R and D loads only but not E, why). This high output state and voltage is frequently denoted by the alternative symbols V_{HH} or V_H. This is low power state which can be used to advantage in designing circuits on a chip to minimize the power dissipation. With the input MOST normally off (in off state at $v_G=0$), a gate voltage step is required to turn it on in order to switch the output to the low-voltage state. If a depletion-mode nMOST or pMOST were used as the input device, an additional d.c. power supply is required to bias either the gate, V_{GG}, or the substrate, V_{XX}, of the input device to cutoff the existing channel current. This increases the design complexity and the manufacturing cost. Nevertheless, the substrate bias was used, known as the **body effect**, to cut off the n-channel on the p-Si induced by positive oxide charge. This enabled IBM to build the first all-transistor mainframe computer by using Si nMOST DRAM memory to replace magnetic core memory. The oxide-charge induced n-channel has been eliminated in recent Si VLSI technology by ion implantation in the Si surface channel, and p+boron or n+phosphorus doped poly-Si gate, which are used also to precisely adjust the threshold voltage to assure that all the MOST's are enhancement mode transistors so that a substrate bias is no longer needed.

Thus, we shall study only the inverter circuits that use the enhancement-mode MOST as the input device. Furthermore, only the first four inverters, Figs.672.1 to 672.4, using the enhancement nMOST as input will be analyzed since the analytical results are also applicable to pMOST shown in Figs.672.6 to 672.9. These are the eight basic digital circuit function blocks from which most of the complex MOS digital integrated circuits are built. The other twelve circuits are less desirable because of the additional d.c. bias for the substrate and the less controllable and hence more expensive fabrication processes that give the built-in channel. Their analyses could be covered by an MOS encyclopedia for completeness rather than practicality.

Simplified cross sectional views of the first four inverters are shown in Figs.672.21(a)-(d) to illustrate their structure in integrated circuit chips. The circuit diagrams shown in the **primed** figures, Figs.672.21(a')-(d'), are direct topological images of the cross-sectional views, while those in the **doubly primed** figures, Figs.672.21(a")-(d"), are the more conventional orientations used in circuit designs.

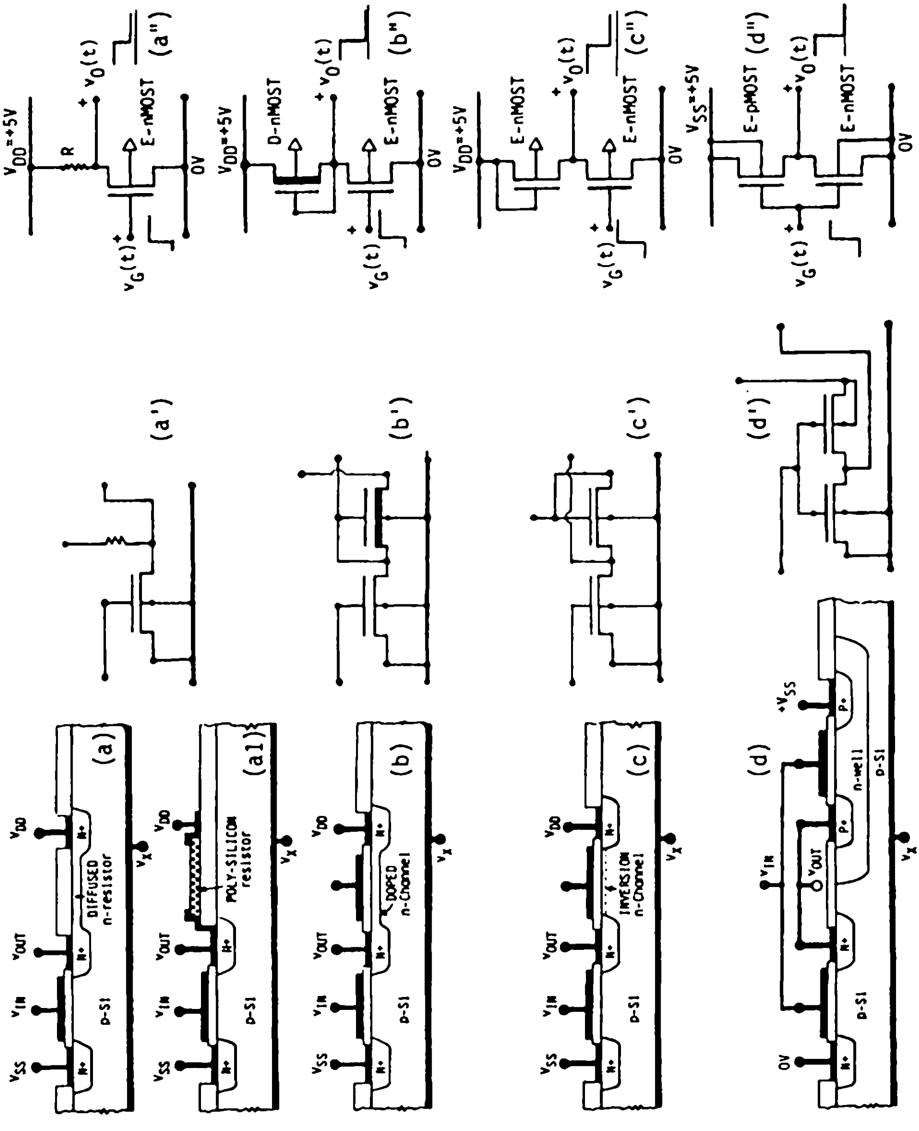

V_DD=+5V
R
+ v_O(t)
E-nMOST
(a")
v_G(t)
0V
D-nMOST
(b")
(c")
V_SS=+5V
E-pMOST
(d")
(a')
(b')
(c')
(d')
V_SS
V_IN
V_OUT
V_DD
V_X
N+
DIFFUSED n-resistor
p-Si
(a)
POLY-SILICON resistor
(a1)
DOPED n-Channel
(b)
INVERSION n-Channel
(c)
0V
+V_SS
P+
n-well
(d)

Fig.672.21 The cross-section view (unprimed) and the circuit diagrams (primed and double-primed) of the four most popular MOS inverters using the nMOST as the input device. **(a)** and **(a1)** RENMOS, the resistance load NMOS inverter where (a) is a diffused n-type resistor and (a1) is a deposited polycrystalline silicon resistor. **(b)** DENMOS, the NMOS inverter with an depletion-mode doped n-channel nMOST load, DENMOS **(c)** EENMOS, the NMOS inverter with an enhancement-mode (inversion or induced n-channel) nMOST load. **(d)** EECMOS, the CMOS inverter with an enhancement-mode pMOST load.

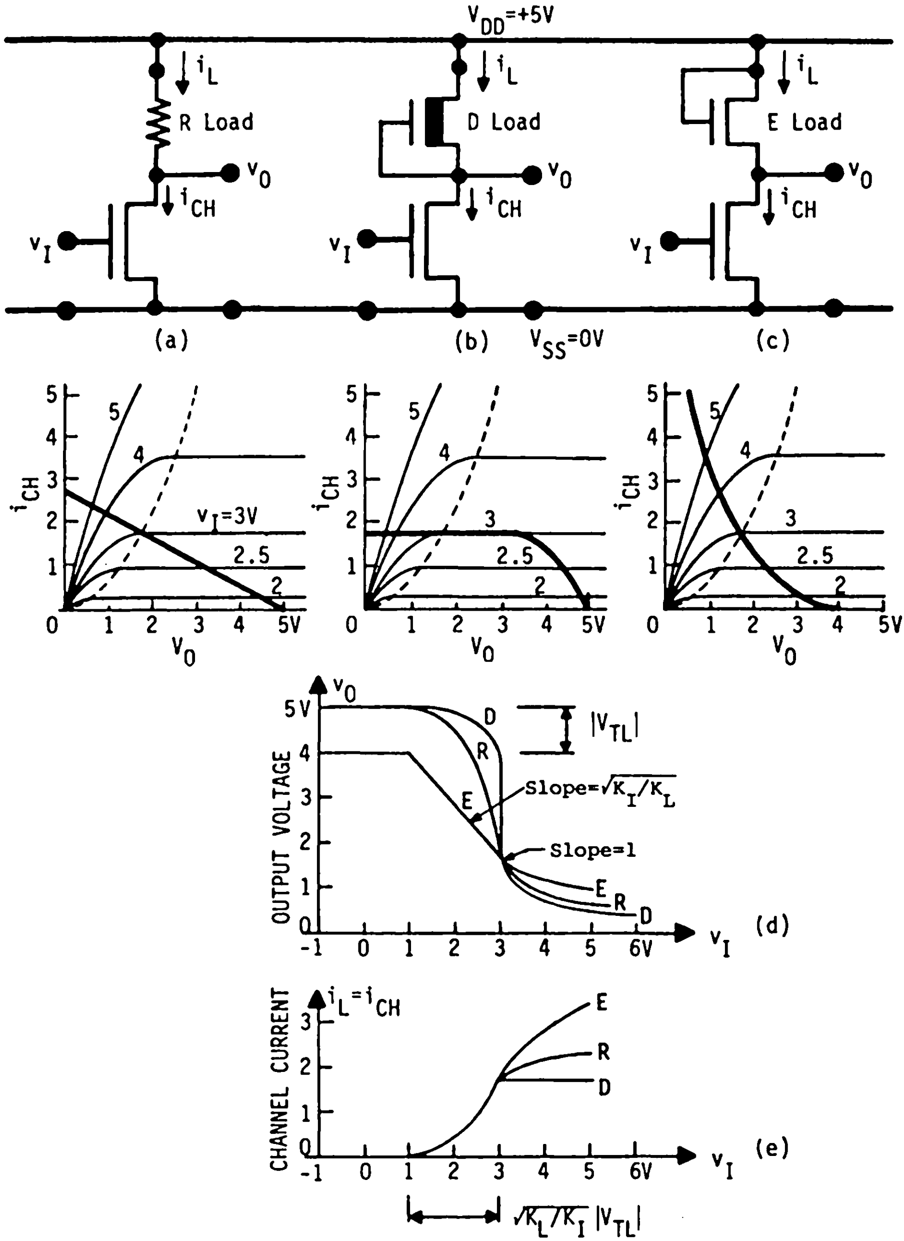

V_DD=+5V
i_L
R Load
D Load
E Load
v_O
i_CH
v_I
(a)
(b)
V_SS=0V
(c)
i_CH
5
4
3
2
1
0
1
2
3
4
5V
v_I=3V
2.5
2
v_O
v_O
5V
4
3
2
1
0
-1
0
1
2
3
4
5
6V
OUTPUT VOLTAGE
D
R
E
|V_TL|
Slope=√K_I/K_L
Slope=1
E
R
D
v_I
(d)
CHANNEL CURRENT
i_L=i_CH
E
R
D
v_I
(e)
√K_L/K_I |V_TL|

Fig.672.22 Characteristics of the three NMOS inverter circuits. The current-voltage characteristics of the input nMOST and the load line of (a) the resistance load RENMOS, (b) the depletion load DENMOS, and (c) the enhancement load EENMOS. (d) The voltage transfer characteristics. (e) The channel current vs output voltage or current transfer characteristics.

Analysis of the Three NMOS Inverter Circuits

The output current-voltage characteristics of the input or drive nMOST and the load lines are shown in Figs.672.22(a)-(c) for the three NMOS inverters, RENMOS, DENMOS, and EENMOS. A threshold voltage of $V_T = 1.0V$ is assumed and the power supply voltage, V_{DD}, is assumed to be +5V. The properties of the load devices are chosen such that the drive or input transistor is out of saturation at an input voltage V_I greater than 3.0V and an output voltage V_O less than 2.0V. These values are chosen in order to facilitate the comparison of the performance of the three inverter circuits shown in parts (d) and (e) of this figure.

The output voltage versus input voltage, known as the **voltage transfer characteristics**, are shown in Fig.672.22(d) where the three curves are: R, resistance load; D, depletion load; and E, enhancement load. It is evident that the depletion load inverter gives the sharpest transition from the HIGH output state to the LOW output state. It also gives the largest output voltage swing and the lowest output voltage at LOW or the lowest standby power dissipation.

Figure 672.22(e) shows the channel current versus the input voltage of the three circuits. These are known as the **current transfer characteristics**. Channel current is an indicator of power dissipation. Again, the depletion load (D-load) gives the lowest current when the drive transistor is in the on-state or the input voltage is high and the output voltage is low. Thus, the D-load circuit should provide the lowest power dissipation.

Figures 672.22(b) and (d) also indicate that the D-load inverter circuit can provide a large current drive during the switching transient. Note the horizontal intersection of the two $V_G = 3$ curves in Fig.672.22(b) that covers the output voltage range of 2 V to 3 V. This large current drive can provide a faster charge/discharge cycle of the load capacitance. The high performance nMOS 16k SRAM (16 kilobit Static Random Access Memory; SRAM is to be discussed in the next section, 673) introduced in 1979 employed such a depletion load nMOST. Resistive load has also been used in high density SRAM during earlier development in 1979. One approach actually used a gate-controlled resistance (actually a MOST) fabricated by depositing polysilicon film using a low pressure chemical vapor deposition technique (LPCVD). This is really a CMOS inverter circuit whose load is an enhancement mode p-channel MOST while its drive is an n-channel MOST. Poly-Si resistance load has been a renewed solution for on-chip SRAM in digital logic chips since 1990. CMOS is discussed in the next subsection.

To obtain the voltage and current transfer characteristics shown in Figs.672.22(d) and (e), intersections of the load line with the family of I_D-V_D of the drive transistor are necessary. These intersections are indicated in Figs.672.22(a)-(c). The simplified current-voltage equation given by (670.2B) and (670.2C) are used to reduce the algebra. The subscript I is used for the input nMOST and L for the load nMOST, such as K_I, V_{TI}, K_L, and V_{TL}. V_I is the input voltage which is also the gate voltage of the input transistor, V_{GI}. V_O is the output voltage which is also the drain voltage of the input transistor, V_{DI}.

In the nonsaturation range when $0 < V_O < V_I - V_{TI} <$

$$I_D = K_I[2(V_G-V_{TI})V_D-V_D^2] = K_I[2(V_I-V_{TI})V_O-V_O^2] \tag{672.1}$$

and in the saturation range when $V_I-V_{TI} < V_O$

$$I_D = K_I[(V_G-V_{TI})^2] = K_I[(V_I-V_{TI})^2]. \tag{672.2}$$

The factor 1/2 is absorbed into K_I in the above equation, $K_I=(W/L)(\mu_nC_o/2)$, in order to simplify the algebraic expressions. These two equations must be solved simultaneously with the I-V equation of the load device. The solutions are given for the three inverters in the following subsections.

The RENMOS Inverter

For the resistance load NMOS inverter given by Fig.672.22(a), the I-V equation for the load is given by

$$I_R = I_D = (V_{DD}-V_O)/R. \tag{672.3}$$

Thus, in the saturation range of the drive transistor, we would solve (672.2) and (672.3) simultaneously by eliminating I_D between them to give

$$I_D = K_I(V_I-V_{TI})^2 = I_R = (V_{DD}-V_O)/R$$

or

$$V_O = V_{DD}- RK_I(V_I-V_{TI})^2. \qquad \textbf{Saturation range} \tag{672.4}$$

This solution gives the inverted half-parabola labeled R in Fig.672.22(d) between $V_I=1$ and 3V.

In the nonsaturation range of the input transistor, we would solve (672.1) and (672.3) simultaneously since (672.1) describes the I-V of the drive transistor in the nonsaturation range. This gives

$$I_D = (V_{DD}-V_O)/R = K_I[2(V_I-V_{TI})V_O - V_O^2] \qquad \textbf{Nonsaturation range}$$

or

$$RK_IV_O^2 - [2RK_I(V_I-V_{TI}) + 1]V_O + V_{DD} = 0 \tag{672.5}$$

which is a quadratic equation of V_O. The solution is given by

$$V_O = \frac{[2RK_I(V_I-V_{TI})+1] \pm \sqrt{[2RK_I(V_I-V_{TI})+1]^2 - 4RK_IV_{DD}}}{2RK_I} \qquad (672.6)$$

The negative sign is to be used. Alternatively, we may write V_I as a function of V_O without having to solve a second order algebraic equation since from (672.5)

$$V_I = V_{TI} + [(V_{DD}-V_O) + RK_IV_O^2]/2RK_IV_O$$

$$= V_{TI} + [(V_{DD}/V_O)-1]/2RK_I + V_O/2. \quad \textbf{Nonsaturation range} \qquad (672.7)$$

The intersection of (672.4) and (672.6), indicated by the point $V_I=3$ in Fig.672.22(d), can be readily obtained from (672.5) since this is the point where $V_I-V_{TI}=V_O$ or the slope is equal to 1. Making this substitution in (672.5), then

$$V_{DD}-V_O = RK_IV_O^2$$

or

$$V_O = (1/2RK_I)[\sqrt{1+4RK_IV_{DD}} - 1] = V_I - V_{TI}. \qquad (672.8)$$

The DENMOS Inverter

For the depletion-mode load device shown in Fig.672.22(b), the two equations for the two I_D-V_D ranges of the load transistor are obtained by letting $V_G=V_S=V_O$ and $V_D=V_{DD}$ in (670.2C). These are

$$I_L = K_L[2(-V_{TL})(V_{DD}-V_O) - (V_{DD}-V_O)^2] \quad \textbf{Nonsaturation range} \qquad (672.9)$$

and

$$I_L = K_L[(-V_{TL})2] \quad \textbf{Saturation range} \qquad (672.10)$$

Note that V_{TL} is negative in the depletion-mode nMOST.

The V_O vs V_I characteristics shown in Fig.672.22(b) must be individually obtained for the three input voltage ranges: V_I from 1 to 3V, $V_I=3$V and $V_I \geq 3$V. This is necessary since in the middle range, both the drive and load transistors are in their drain current saturation range while in the other two ranges one transistor is in saturation and the other is not. The solutions, consisting of three sets of simultaneous algebraic equations, are given by (672.11)-(672.13). They can then be solved numerically for a given set of K_I, K_L, V_{TI}, V_{TL}, and V_{DD}. The solutions, presented as the voltage and current transfer curves (V_O-V_I and I_D-V_I), are shown in Figs.672.21(d) and (e) with the label D. These curves show that among the REnMOS, DEnMOS and EEnMOS inverters, the DEnMOS inverter has the largest output voltage swing, the sharpest voltage transition and hence the largest noise margin, and the smallest standby current or power dissipation.

(1) INPUT EnMOST IN SATURATION
LOAD DnMOST IN NONSTUARATION

$$I_I = (672.2) = K_I(V_I-V_{TI})^2$$

$$= I_L = (672.9) = K_L[2(-V_{TL})(V_{DD}-V_O)-(V_{DD}-V_O)^2] \qquad (672.11)$$

(2) INPUT EnMOST IN SATURATION
LOAD DnMOST IN SATURATION

$$I_I = (672.2) = K_I(V_I-V_{TI})^2 =$$

$$= I_L = (672.10) = K_L(-V_{TL})^2$$

or

$$V_I = V_{TI} + \sqrt{K_L/K_I}(-V_{TL}) \equiv V_{TR} \qquad (672.12)$$

This input voltage is known as the **transition voltage**, V_{TR}, and it is labeled on Figs.672.21(d) and (e). One can also show that the initial drop of the output voltage from V_{DD} is $-V_{TL}$ as indicated on Fig.672.22(d).

(3) INPUT EnMOST IN NONSATURATION
LOAD DnMOST IN SATURATION

$$I_I = (672.1) = K_I[2(V_I-V_{TI})V_O-V_O^2]$$

$$= I_L = (672.10) = K_L(-V_{TI})^2 \qquad (672.13)$$

The EENMOS Inverter

A similar algebraic procedure can be used to obtain the solutions for the EENMOS inverter using an enhancement-mode nMOST as the load shown in Fig.672.22(c). The E-nMOST load is always in saturation making the algebra simpler. The V_O-V_I and I_L-V_I transfer characteristics are labeled E in Figs.672.22(d) and (e). It is obvious that among the three nMOS inverters, the EEnMOS inverter is the worst in all three properties: smallest output swing, most gradual transition and hence smallest noise margin, and highest power dissipation. But, it is the easiest to fabricate and hence the most widely used in circuits.

Power Dissipation of the Three NMOS Inverters

The three nMOSFET inverter circuits just described, RENMOS, DENMOS and EENMOS, dissipate little power when the **input is low** and **output is high** since the channel current is essentially zero, consisting of just the drain junction leakage current of the input transistor, as indicated by the left portion of the V_O-V_I and I_{CH}-V_I curves in Figs.672.22(d) and (e). However, the power dissipation is significant when the **input is high** and **output is low**, because the channel current is high as indicated by the right portion of these two figures. The significant power dissipation comes from the load transistor when the output is low. It is equal to the sum of the product $I_L \cdot V_{DD}$ of all the conducting load transistors on a chip.

In a gate array chip, about half of the transistors are in the on-state and half in the off-state under a typical circuit operation condition. Thus, the average standby power dissipation in a D-NMOS chip is about 50% of the sum of $I_{DS} \cdot V_{DD}$ of each of the nMOS transistors on the chip. The power dissipation of a gate array chip containing 10k gates or 10k DE-NMOS inverters can be estimated using the numerical calculation made in section 644 which gave $I_{DS}=250\mu A$, $g_{ms}=1mS$ and $C_o=0.36pF$ for a one-micron ($Z=L=1\mu m$, $x_o=100A$) state-of-the-art nMOS transistor. Thus, the power dissipation is

$$P_{standby} = I_{DS} x V_{DD} x 10^4 x (1/2)$$

$$= 250x10^{-6}x5x10^4x0.5 = 6.25 \text{ Watts.} \qquad (672.14)$$

This is considerably higher than the rated nominal heat dissipation rate of a dual-in-line plastic package, about 500mW to 1W. The lower power, ~1W, is necessary in order to limit the chip temperature to less than the maximum temperature of operation, 85C, in order to assure a 10-year operating life. The power dissipation increases proportionally in larger gate arrays. Current technology has increased the gate counts to well above 100k. The microprocessor chips for the personal computer, INTEL's 80486 (1990 introduction cost ≈ $1000/unit) and Motorola's 68040, have over one million transistors, most of which are logic gates. Thus, in large and dense gate array chips, ceramic package with good heat sinks is required. Consequently, package cost is a substantial fraction and frequently more than 50% of the cost of a million-transistor CPU or logic chip. High rate of power dissipation are also be designed into the system box to remove the heat generated from the high power density in each chip, such as using a blower for convective cooling.

The CMOS Inverter Circuit

We shall first provide some general background on the CMOS inverter circuit. Then, a d.c. circuit analysis of the output and transfer characteristics and a simple estimate of the power dissipation will be described.

Why CMOS?

Because of the large power dissipation of the three N-MOS inverter circuits, there was a strong motivation to find a circuit which reduces the power dissipation. This is the fourth MOS inverter circuit, the CMOS (Complementary MOS), whose input is an enhancement-mode nMOST and whose load is an enhancement-mode pMOST. Thus, when the inverter is not switching, one of the two transistors is non-conducting. Hence, there is no standby or quiescent channel current. The stand-by power dissipation is all due to the junction leakage current which is in the nano- or pico-watt range at +5V depending on the area of the junctions. **Power dissipation or energy consumption occurs only during switching transient.**

Although energy is consumed only during switching, the power dissipation is proportional to the frequency of switching, $P = \text{Capacitance} \cdot V_{DD}^2 f_{clock}$. Thus, it has often been argued by textbook authors and practicing engineers that (i) when CMOS is run as fast as NMOS, there is no power dissipation advantage for CMOS, and (ii) special power-down circuits can be designed into the NMOS chip so that the NMOS dissipates little power while not switching. The former is indeed the case only for circuits operated at very high clock frequencies. The latter requires significant peripheral control circuits to power down the chip and even at power down, the dissipation can still be 1000 times higher than the nanowatt or picowatt standby power dissipated by the CMOS chip. Thus, these are not general results as insisted by some. Everyday examples are the one-year or longer operating life of the battery operated digital wrist watch and hand-held electronic calculator, whose longevity owes to CMOS's low power dissipation.

There is a second distinct CMOS advantage over NMOS because one of the two transistors is in the off-state. Since one of the two transistors is off, the output voltage is either at the ground potential, 0V, or at the power supply voltage, +5V, so that **the output voltage swing is equal to the full power supply voltage.** This larger voltage swing gives higher noise immunity and larger noise margin.

These two CMOS advantages will be further elaborated and quantitatively demonstrated using analytical solutions to be derived in the following paragraphs. CMOS may **latch-up** into a high current low voltage state by random noise. Its physical origin and solution will be described.

Evolution History of CMOS Structures

Among the eight CMOS circuits shown in Figs.672.1-20, the most commonly used is the EECMOS shown in Figs.672.4 or 672.9, first suggested by Wanlass during 1962-3. A modern monolithic design is shown in Fig.672.21(d) which consists of a n-channel MOST (nMOST) on a p-Si substrate and a p-channel MOST (pMOST) in the n-well on the p-Si substrate. In order to operate properly (zero standby power and minimum number of biasing voltages), the nMOST and pMOST must both be enhancement mode MOSTs. Historically, the enhancement mode nMOST was difficult to fabricate compared with the enhancement pMOST because the positive threshold voltage ($V_T > 0$) of the enhancement mode nMOST was very hard to reproduce even in the laboratory due to the uncontrollable residual positive charge in the gate oxide at the time. Precise manufacturing control to give a designed positive threshold voltage for the nMOST using the **bulk-charge body effect** analyzed in 1964 by this author, rather than the **substrate-bias body effect** invented by Heiman at RCA in 1966, was not successful in 1965 because of large random variation of resistivity or dopant impurity concentration over a 3" wafer and from wafer to wafer. Substrate surface doping control of threshold voltage was not attainable until the ion implantation technology for precision doping was developed and ion accelerators for semiconductor manufacturing became available during 1969-1970. CMOS circuits produced by implantation were first reported by Si transistor technologists at Hughes Aircraft Company (1968) and by a Mostek-Sprague collaboration team in 1970 (J.MacDougall, K.Manchester, and R.B.Palmer, Electronics, June 22, 1970). Ion implantation has been the only production technique to this day (1991) that can reproducibly incorporate a minute amount of boron acceptor (or phosphorus-arsenic donor) into the surface layer of a large area Si wafer (6"-8"). The total dose of boron acceptor impurity required to control the threshold voltage of nMOST is in the 10^{11} boron/cm^2 range. This amount can adjust the threshold voltage by 500mV in a 1000A gate oxide or 50mV in a 100A gate oxide. This is indeed a very minute amount of impurity atoms: it is only one-thousandth of a monolayer of boron atoms on the silicon surface. Such a low amount cannot be reproducibly deposited by other means such as diffusion, epitaxial growth, or evaporation. Soon afterwards (~1972) CMOS chips were manufactured in volume as the frequency divider for digital wrist watch using ion implantation. CMOS was necessary in order to attain a 1-year battery life of the wrist watch. A 1980 multi-function two time zone CMOS wrist watch (TIMEX-quartz) uses a 700-Joule 1.5V silver-oxide battery to give more than one year life. A CMOS programmable handheld scientific calculator of about 100 built-in functions (HP-11C) uses three 1.5V silver oxide batteries and can run a program continuously for about 180 hours. High performance 32-bit workstations in excess of 10Mips (Mips=million instructions per second) and 10MFlops (MFlops=million floating point operation per second) are now manufactured using CMOS with only air cooling instead of water or liquid cooling. Latest entries (1991) are approaching 100MiPs selling at about $20k: the RISC workstations IBM320C (29.5Mips, 8.5MFlops), DEC5000PX (24Mips,3.7MFlops), HP730CRX (76Mips,22MFlops) and Intel's RISC chip I860XP (50Mips,>100MFlops); and Intel's CICS chip I486DX (50Mips) and future scaled down versions, Ix86 ($x \geq 5$, >100Mips). All of these advances in solid-state and computer electronics have taken advantage of the nearly zero standby power dissipation of CMOS. A review of the history of invention and recent applications of CMOS was given in October 1988 [600.1].

A serious electrical instability can occur in the 1-well CMOS shown in Fig.672.21(d) causing it to **latch-up** to a low-voltage/high-current (LVHI) state during operation when the cell dimension is shrunk. The latch-up is due to the presence of parasitic lateral n+/p/n/p+ four-layer transistor action along two paths (the interfacial and bulk paths) between the left nMOST and right pMOST in Fig.671.21(d). The two minority carrier pathways are labeled in Fig.672.23(b) and (c). The four-layer device acts as a n+/p/n bipolar junction transistor (BJT's are studied in chapter 7) regeneratively coupled to a p/n/p+ bipolar junction transistor. The regenerative feedback gives the **bistable** states: a high-voltage/low-current

(HVLI) normal state and a low-voltage/high-current (LVHI) **latch-up state.** The CMOS can be triggered into the latch-up state by a transient noise voltage that appears in one of the CMOS terminals but it cannot be easily switched back to the normal HVLI state by a control signal voltage. Usually, the d.c. voltage must be turned off to restore it to the normal HVLI state. Thus, once latched up into the high-current state, the CMOS circuit ceases to function and the high current may even destroy the the CMOS chips due to heating.

Silicon technology advances in the mid-1980's have produced ingenious CMOS cell designs which have essentially eliminated the latch-up instability, but at a significant increase of manufacturing cost due to more critical processing steps which lower yields. The evolution of latch-up immunity engineering is shown in Figs.672.23(a) to (c) which led to the latch-up immune epitaxial twin-well structure of figure (c). The history is now briefly described using these figures.

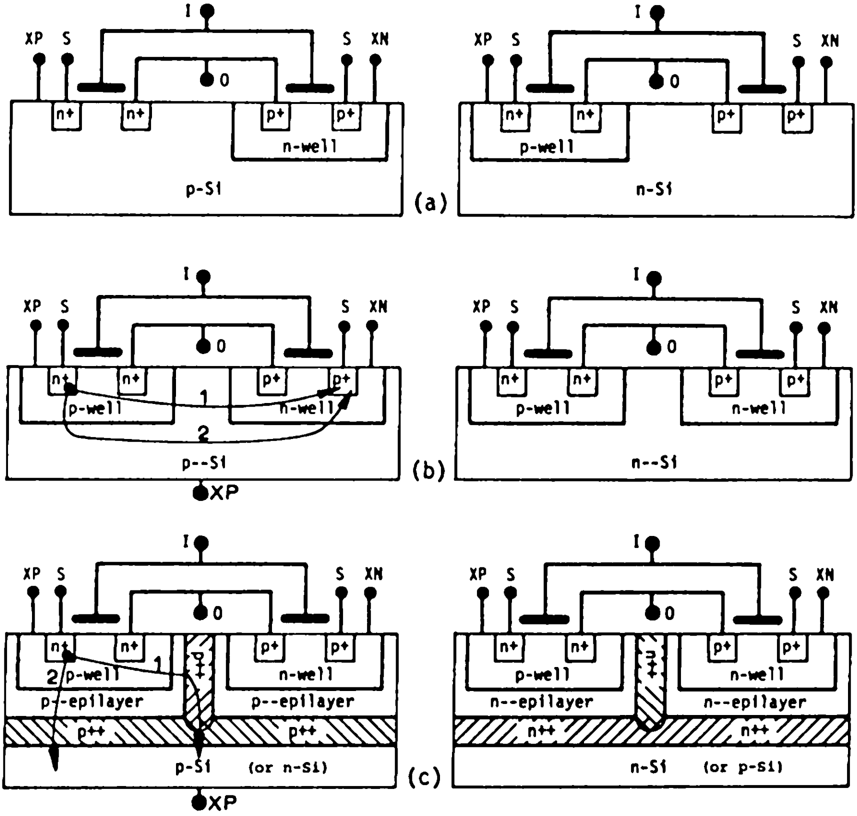


Fig.672.23 Cross sectional view of physical realizations of the CMOS inverter. **(a)** One-well and **(b)** two-well on bulk (non-epitaxy) Si substrate. **(c)** Two-well on epitaxial substrate with latchup prevention via degenerate low/high (--/++) junction barrier (shaded).

Figure 672.23(a) shows the two original CMOS designs using one well. The left figure has a n-well on a p-Si substrate. The right figure has a p-well on a n-Si substrate. The well is made by ion implantation through an oxide window. Ion implantation is a highly reproducible and controllable manufacturing step which can readily give precise junction depth or well thickness and precise implanted impurity concentration in the well. Thus, the variation of the threshold voltage of the MOSTs in the well among all the in-well transistors on a 6"-8" Si wafer should be quite small. However, the second MOST in the CMOS is on the substrate and not in an ion-implanted well, thus, it will still suffer excessive threshold voltage variation over the 6"-8" Si wafer, but also from wafer to wafer. This variation comes from the large tolerance usually specified for the starting resistivity of the 100 or more 6"-8" Si wafers used in a VLSI chip production run. A VLSI chip containing 100k CMOS inverter gates will probably not function because of the large threshold voltage variation among the 20k to 200k inverter transistors. To circumvent, the twin-well structures shown in Fig.672.23(b) were invented by Bell Labs. The two wells are both ion implanted on a common high-resistivity (>50Ω-cm) p- or n-Si substrate. Thus, the detrimental bulk-charge threshold voltage variation, from the large spread in substrate dopant impurity concentration, is eliminated because the impurity concentration on the surface of the two wells can be precisely controlled by implantation of acceptor (B) and donor (P,As,Sb). The high resistivity p^{--}substrate on the left of figure (b) has been the production preference for both NMOS and 2-well CMOS because uniformly doped, 6"-8" diameter, 2'-long, and >50Ω-cm p-type Si crystal is easier and less expensive to grow than n-type. This preference arises because boron is the residual impurity in Si owing to its near unity (k=0.8) distribution or segregation coefficient between Si liquid and solid while the other impurities have lower value: P(k=0.35), As(k=0.3) and Sb(k=0.023). Crystal growth is described in section 134.

To increase the latch-up voltage in CMOS so that circuit noise voltage transients cannot initiate a latch-up, two degenerate layers are incorporated into the 2-well CMOS cell of Fig.672.23(b) to reduce regenerative feedback. They are shown in figure (c) as shaded layers. Consider the CMOS on the left. It is seen that they are inserted into the p-base region of the parasitic n+/p/n-well bipolar transistor. Their presence prevents or substantially reduces the minority carrier (electrons) current flowing through the surface and bulk feedback paths labeled by arrow-1 and arrow-2 respectively arrows in figures (b) and (c). Note that the degenerate layers give the n+/p/--**|degenerate-barrier|**--/n/p+ structures where -- is either p-- in the left CMOS or n-- in the right CMOS. The degenerate p^{++} or n^{++} barrier reduces the bipolar current gain via the two positive feedback pathways of the bipolar transistor with the well-collector to such a low value that the high current latch-up condition cannot be reached by a noise voltage as high as or even greater than the power supply voltage. Consider again the CMOS cell with the p--epilayer Si substrate on the left of Fig.672.23(c). The starting p-Si wafer has a noncritical (inexpensive) medium resistivity. It has a noncritical low-resistivity (0.01 to 0.001Ω-cm) p^{++}layer from ion implant or epitaxy. The p^{++}layer is covered by a thin (a few μm) high-resistivity (0.1-10Ω-cm) and high quality epitaxial p^{--}layer. A vertical p++ isolation wall is then formed by boron diffusion or implantation. Figure 672.23(c) shows that the two minority carrier (electrons) diffusion paths (arrow-1 and arrow-2) are now blocked by the potential barrier from the p--/p++ junction of the p^{++} isolation wall and the p^{++} epitaxy layer. Thus, few electrons can reach and be collected by the p+/n-well source junction of the pMOST, resulting in very low current gain so that latch-up will not occur. Most of the electrons are directed to the substrate contact.

Although the p/n/p/n bistability and latchup is undesirable in CMOS, it is in fact the principal mechanism of operation of the silicon controlled rectifiers (SCR) described in sections 78n. SCR's have dominated the power control applications over a tremendously wide range of power levels, from 60-watt light dimmer to kilowatt electrical heaters and motors, and even megawatt electrical locomotives, using all solid-state electronic control systems that contain no power vacuum tubes.

D.C. Analysis of the CMOS Inverter Circuit

The circuit, load line, and transfer characteristics of the CMOS inverter are given in Figs.672.24(a)-(d). The mathematical analyses to give the transfer characteristics are similar to those of the three NMOS inverters. The solutions are listed below. Since both the nMOST and the pMOST are enhancement devices, we have $V_{TN}>0$ and $V_{TP}<0$ in these solutions.

(A) The CMOS Equations

1. $V_O = V_{SS}$ and $I_S = 0$ (672.15)
2. $V_O = V_I - V_{TP} + \sqrt{[V_{SS}-(V_I-V_{TP})]^2-(K_N/K_P)(V_I-V_{TN})^2}$ (672.16)
   $I_S = K_N(V_I-V_{TN})^2$ (672.17)
3. $V_{TR} = V_I = [V_{SS}+V_{TP}+\sqrt{K_N/K_P}V_{TN}]/[1+ K_N/K_P]$ (672.18)
   $I_{TR} = [K_N K_P/2(\sqrt{K_N}+\sqrt{K_P})^2](V_{SS}+V_{TP}-V_{TN})^2$ (672.19)
4. $V_O = V_I - V_{TN} - \sqrt{(V_I-V_{TN})^2-(K_P/K_N)[V_{SS}-(V_I-V_{TP})]^2}$ (672.20)
   $I_S = K_P[V_{SS} - (V_I - V_{TP})]^2$ (672.21)
5. $V_O = 0$ and $I_S = 0$ (672.22)

(B) Power Dissipation

Power dissipation of the CMOS arises entirely from the dynamic switching energy dissipation since there is no standby power dissipation. This is also obvious from an inspection of the V_O-V_I and I_{CH}-V_I solutions given in Figs.672.24(c) and (d). The switching power dissipation can be estimated using these figures which show that the power dissipation during one transition from output high to output low or vice versa is approximately

$$V_{SS}\cdot(I_{TR}/2)\cdot2\cdot(t_d\cdot f_{clock}) = V_{SS}\cdot(I_{DS}/8)\cdot(t_d\cdot f_{clock}) \qquad (672.23)$$

where we have assumed that the two transistors are identical and their threshold voltages are nearly zero. Larger threshold voltages would give smaller current, lower transient power dissipation, and larger noise margin. Thus, for $V_{SS}=5V$ and $I_{DS}=250\mu A$ from the 1-micron transistor, a 10k gate array chip switching 10% of the time ($t_d\cdot f_{clock}=0.1$) would have a power dissipation of

$$P_{switching} = (V_{SS}\cdot I_{DS}/8)\cdot(t_d\cdot f_{clock})\cdot(10k/2) \qquad (672.24)$$

$$= (5\text{x}250\text{x}10^{-6}/8)\text{x}0.1\text{x}5000 = 0.625 \text{ Watt.} \qquad (672.25)$$

This is ten times lower than that just estimated in (672.14) for the 10k gate array using NMOS inverters. If the gates are switched near 100% of the time, then the switching power is 6.25 Watts which equals that using the NMOS inverters.

The above estimate does not take into account the capacitance loading which would significantly increase the gate delay, t_d, and hence the switching duty cycle, ($t_d\cdot f_{clock}$), at a given clock frequency. An alternative estimate assuming that the power dissipation is entirely caused by the transistor dissipation during charging and discharging the load capacitances would give

$$P_{switching} = 2\cdot[(\text{gate-count})/2]\cdot C_L\cdot V_{SS}^2\cdot f_{clock}$$

$$= C_L\cdot V_{SS}^2\cdot f_{clock}\cdot(\text{gate-count}) \qquad (672.26)$$

where C_L is the total load capacitance including the drain junction, the gate, the overlap, the interconnect lead and other parasitic capacitances. With typical values of $C_L=100fF$, $f_{clock}=100MHz$ and $V_{SS}=5V$, the switching power dissipation for 10^4 gates would be

$$P_{switching} = 100\text{x}10^{-15}\text{x}5^2\text{x}10^8\text{x}10^4 = 2.5 \text{ Watts.} \qquad (672.27)$$

This is 4 times higher than the internal power dissipation computed in (672.25) which assumed no circuit loading but at 10% of duty cycle but 2.5 times smaller than switching at nearly 100% duty cycle. In practice, both intrinsic delay and capacitance loading must be considered since logic gate arrays have fan-outs of less than about five so that interconnect line capacitance does not dominate.

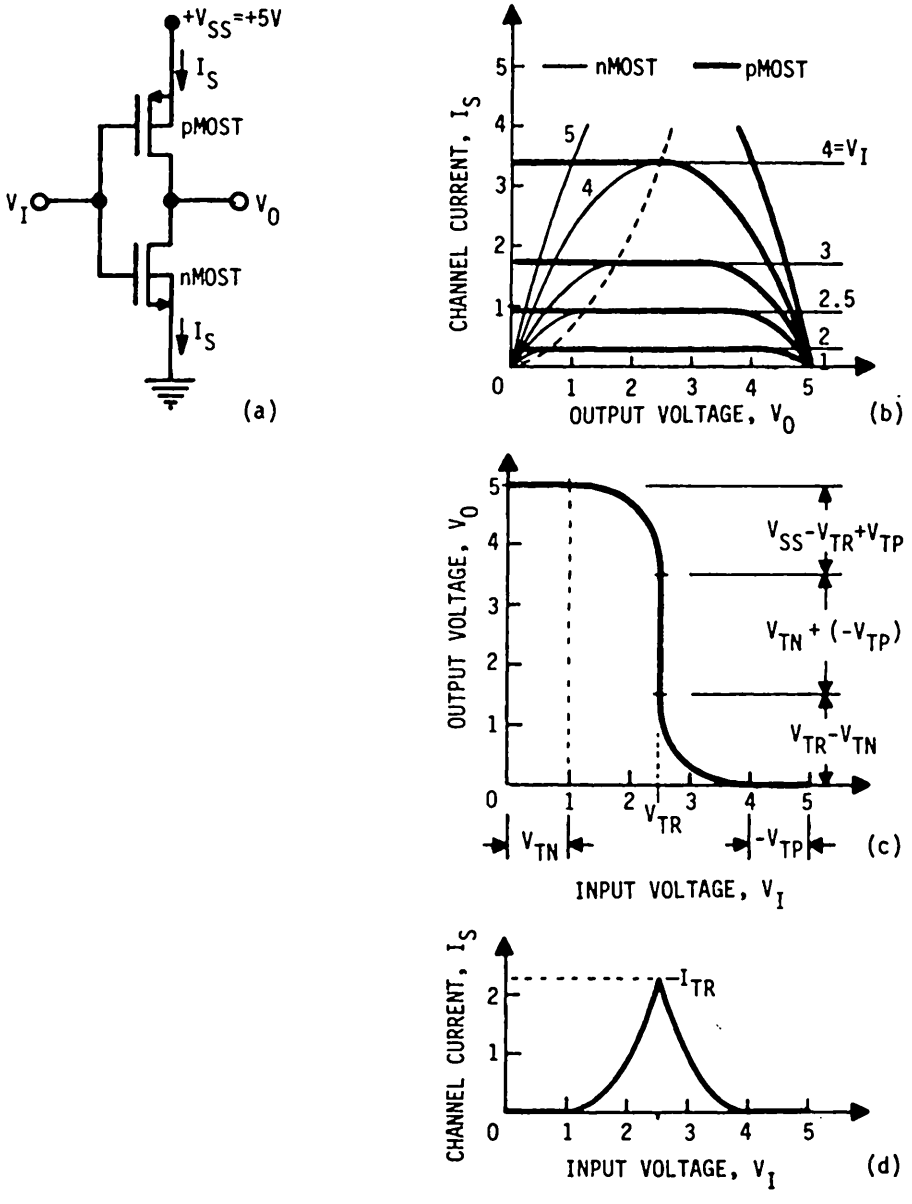

+V_SS=+5V
I_S
pMOST
V_I
V_O
nMOST
I_S
(a)
CHANNEL CURRENT, I_S
— nMOST
— pMOST
5
4
4=V_I
3
2.5
2
1
0 1 2 3 4 5
OUTPUT VOLTAGE, V_O
(b)
OUTPUT VOLTAGE, V_O
V_SS-V_TR+V_TP
V_TN + (-V_TP)
V_TR-V_TN
V_TR
V_TN
-V_TP
0 1 2 3 4 5
INPUT VOLTAGE, V_I
(c)
CHANNEL CURRENT, I_S
I_TR
0 1 2 3 4 5
INPUT VOLTAGE, V_I
(d)

Fig.672.24 The CMOS inverter. (a) The circuit. (b) The I_D-V_D and load line characteristics. (c) The voltage transfer characteristics. (d) The current transfer characteristics.

673 The Static Random Access Memory (SRAM)

The static random access memory (SRAM) cell contains two cascaded inverters whose output is connected to the other input to give regenerative feedback that produces the **bistable states**. Figure 673.1(a) shows the unity feedback loop of two cascaded EE-NMOS inverters of a semiconductor SRAM cell. Figure 673.1(b) gives the symmetrical topology which shows explicitly the cross-coupling of the two inverters. It is called a **flip-flop** circuit since the node voltage can flip-flop between two states, state-0 (0V) and state-1 (+5V). It is also known as the **latch** since the node voltage will latch onto or **memorize** the 1-state if the command pulse is in the 1-state, and the 0-state, if the command pulse is in the 0-state.

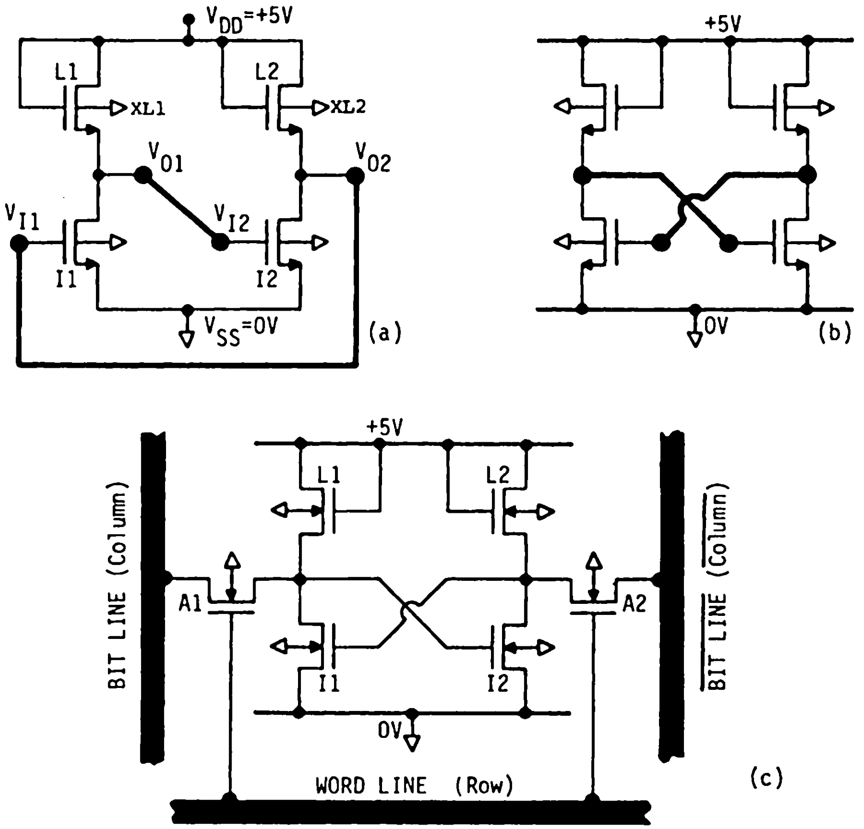


Fig.673.1 Pedagogical circuit development of the SRAM using two cross-coupled EENMOS inverters. (a) The cascade illustration of the 4-T (4-Transistor) bit store showing the unity feedback loop. (b) The symmetrical cross-couple illustration of the bit store. (c) The 6-T SRAM cell containing: 2 access transistors, A1 and A2, 2 bit lines, and 1 word line.

It is **static** because the latched state is retained indefinitely without using a refresh signal as long as the power supply voltage, V_{DD}, is not interrupted. It is **volatile** because it loses the stored state if the power supply is disconnected.

To demonstrate the data retention (latch) and the flip-flop properties, let us use the cascade circuit of Fig.673.1(a). Assume that the initial steady or quiescent state is $V_{I1}=V_{O2}=0V$ and $V_{O1}=V_{I2}=+5V$, and that $V_{DD}=+5V$ and is not interrupted. Suppose that at $t=0$, V_{I1} is connected to a +5V voltage source for 1 second. Then, V_{I2} is switched from +5V to 0V and will stay at 0V indefinitely, i.e. static, after the +5V voltage source is removed from V_{I1}. The reason is as follows. When +5V is applied to V_{I1}, transistor I1 is turned on which pulls down its output to ground, $V_{O1}\rightarrow 0V$. Since $V_{I2}=V_{O1}$, transistor I2 is turned off by $V_{I2}=0$, raising its output towards $V_{DD}=+5V$ which is also the applied input voltage since $V_{I1}=V_{O2}$. Thus, removal of the +5V input voltage source from V_{I1} after one second will not reduce V_{I1} to zero because it is now maintained by V_{DD} through the voltage drop of transistor L2.

A practical SRAM cell contains six MOST's (6-T) is shown in Fig.673.1(c). The two additional MOST's (A1 and A2) serve as **access gate** in the two data bit lines to select a particular cell on a multi-cell SRAM chip. The two gates of the two access MOST's are connected to one word-line (WL) or row, R, while the two drains are connected to two complementary data bit-lines (BLs) or columns, C and $\bar{C}$. From the operation principle just described, it is evident that only one access transistor and one bit line is needed to write and read one SRAM cell. However, SRAM chips using the 5-T and 1-BL cell have not been operated successfully at the maximum designed speed due to statistical variations in the transistor characteristics normally encountered in production which prevent all cells connected to one word line to flip-and-flop.

There are four possible basic SRAM cell structures from the four different loads discussed in the preceding section, 672, (RMOS, EEMOS, DEMOS or CMOS). The historical first monolithic semiconductor memory chip (built by Schmid at Fairchild in 1965) was a 64-bit Si SRAM using the EE-PMOS inverter in its 6-T cells. EE-NMOS took over to give higher speed because $\mu_n>\mu_p$ after eliminating the induced n-channel by substrate bias and later by implanting boron and doping the Si gate to adjust the threshold voltage. The memory cell evolution started from the 6-T EE-NMOS cell and took three paths. The four SRAM cells evolved in two paths shown in Figs.673.2(a)-(d). The third path led to the three DRAM cell structures (all invented by IBM researchers) shown in Figs.673.3(a)-(d).

In the first SRAM evolution path, the EENMOS inverter in Fig.673.2(a) was replaced by the DENMOS inverter in Fig.673.2(b) to give higher speed. In order to reduce standby current ($<1\mu A$ per cell) and power ($<100\mu W$ per chip) and increase output voltage, the two depletion load transistors were replaced by two deposited polycrystalline silicon resistors, resulting in the 6T RNMOS cell in

Fig.673.2(c). The interconnect topology and the fabrication are also simplified. However, low standby current requires a high load resistance which is difficult to reproduce and more expensive to manufacture.

In the second SRAM evolution path, the EENMOS inverters in Fig.673.2(a) were replaced by the CMOS inverters in Fig.673.2(d). Standby power is reduced to the nanowatt level. Output is raised to rail-to-rail (0V and 5V) as indicated by Fig.672.24(c). Speed is not sacrificed by the lower hole mobility in the two pMOSTs because the p-channel is made shorter than the n-channel.

The third evolution path gives the three DRAM cells shown in Figs.673.3(a)-(d) all invented by IBM researchers. The 6-T SRAM in Fig.673.3(a) can also be considered as a charge storage cell with information stored on the two large capacitors connected to the two inputs. Thus, it is a 6-T/2-C cell. The capacitance charge or voltage is maintained by the load transistor or resistor so that they do not decrease with time due to leakage (static). However, the information can still be stored on the capacitors if the two load transistors or resistors are removed, provided the cell is periodically refreshed (dynamic) to replenish the charge lost through the leakage paths which are the resistance of the drain and source p/n junctions and the surface space-charge layer. Removal of the two loads gives the 4-T/2-C DRAM cell shown in Fig.673.3(b), consisting of a 2-T/2-C store and two transistor access gates invented by Spampinato and Terman of IBM in 1970. It is immediately obvious that half of the 2-T/2-C cell is not only redundant but also bit-destructive during read. Thus, it can be discarded, giving the 3-T/1-C DRAM cell shown in Fig.673.3(c) invented by Palfi of IBM in 1971, consisting of the 1-T/1-C store and 2-T gate for nondestructive read . This 3-T DRAM cell was used by Intel in 1970 to produce the first 1-kbit Si DRAM chip.

The 3-T/1-C DRAM cell was further simplified by executing the read also from the write line which eliminates the 2-T gate of nondestructive read. This gave the 1-T/1-C Dennard DRAM cell shown in Fig.673.3(d). However, the read is now destructive. Hence, a restore or write is necessary after read. In spite of the additional restore circuits, the transistor count on a DRAM chip is still smaller using the destructive-read Dennard cell than the nondestructive-read Palfi cell since one restore circuit can be used for a group of Dennard cells while each Palfi cell (3-T/1-C) has two more transistors than the Dennard cell (1-T/1-C).

Although the 1-T/1-C Dennard DRAM cell occupies much less area than the 6-T CMOS SRAM cell, the DRAM cell is slower because it requires a large capacitance to store the bit in order to give a sufficient noise margin while the SRAM bit voltage is maintained by a load transistor or resistor on a very small node capacitance with essentially complete noise immunity. Larger capacitance slows down the access speed. Thus, 1-T DRAM is used for high-density (>Mbit to Gbit) medium-speed memory while 6-T SRAM is used for medium-density (~MBit to <100MBit) high-speed memory.

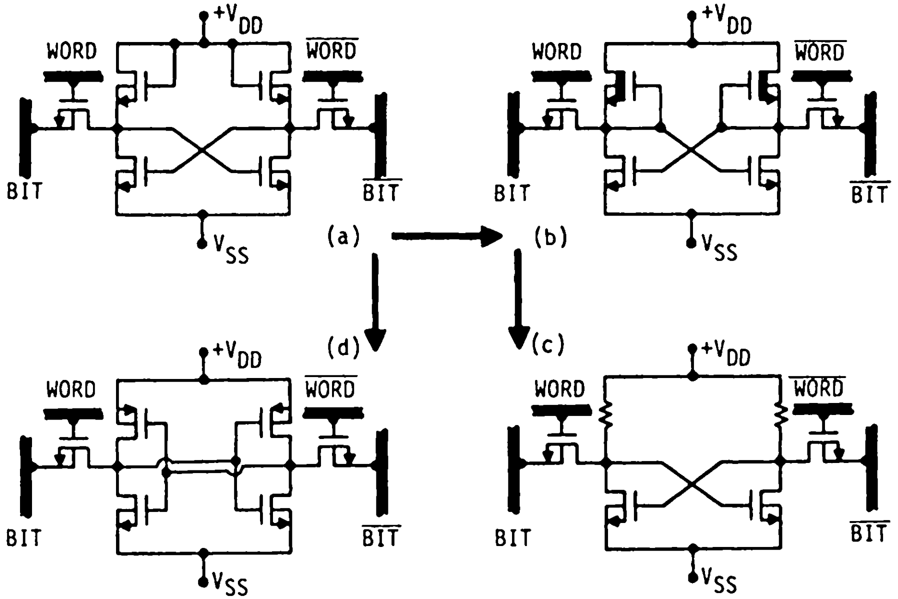

+V$_{DD}$
WORD
$\overline{\text{WORD}}$
BIT
$\overline{\text{BIT}}$
V$_{SS}$
(a)
(b)
(c)
(d)

Fig.673.2 The two evolution paths of the four SRAM cells using the four types of MOS inverters. (a) EE-NMOS inverters. (b) DE-NMOS inverters. (c) R-NMOS inverters. (d) CMOS inverters.

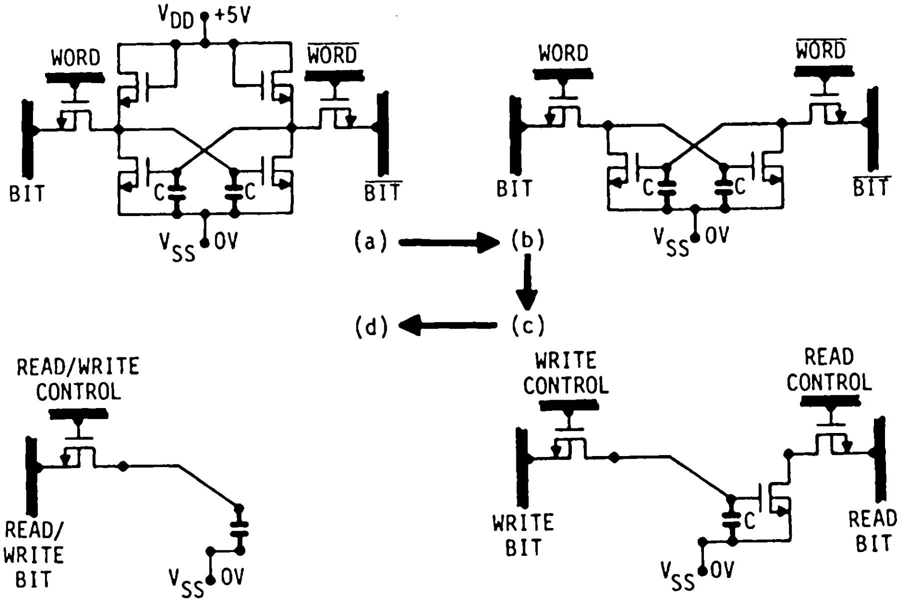

V$_{DD}$ +5V
WORD
$\overline{\text{WORD}}$
BIT
$\overline{\text{BIT}}$
C
V$_{SS}$ 0V
(a)
(b)
(c)
(d)
READ/WRITE CONTROL
READ/ WRITE BIT
WRITE CONTROL
READ CONTROL
WRITE BIT
READ BIT

Fig.673.3 The evolution paths of the four capacitance charge storage DRAM cells. (a) The 6-T DRAM (also SRAM) cell. (b) The 4-T DRAM cell with destructive read (D.P.Spampinato and L.M.Terman 1970). (c) The 3-T DRAM cell with non-destructive read (T.L.Palfi 1971). (d) The 1-T DRAM cell with destructive read (R.H.Dennard 1968).

674 Nonvolatile Random Access MOS Memories (ROM,PROM's)

Many nonvolatile memory (NVM) cells have been invented and mass produced to store repeatedly-used programs and numbers such as the BIOS (Basic Input Output System) code in the IBM personal computers, the phone numbers in memory telephones, and the channel numbers and volume setting in remote-control TV and stereo receivers. The main function of the nonvolatile memory is to store information (a program or a set of numbers) permanently (>10 year) which can be retrieved randomly, rapidly ($<\sim 0.1\mu$s), and repeatedly, on demand, while the stored information needs to be changed or updated only occasionally ($>$daily). A writable-erasable NVM (the Si MOS flash memory) appears destined as the hard-disk replacement in light-weight portable and pocket personal computers whose production has begun to rise sharply in 1990 because of the maturity of several low-power, light-weight technologies (high quality LCD display, CMOS, and high-energy battery). It is ($<\sim 100\mu$s) also much faster than (>100 times) the rotating hard disk (10ms). This mini-**archival** application will greatly increase its market volume and drive down its cost ($<<$\$20/Megabyte). The preceding examples show that the available type and cost of the semiconductor nonvolatile memory are determined and driven by the application-function and market-volume demands.

There are two classes of semiconductor NVM cells (and chips): (i) ROM (Read Only Memory) whose stored information cannot be altered, and (ii) PROM (Programmable Read Only Memory) whose stored information can be changed once or many times (the **EPROM**, Erasable PROM), albeit rather slowly (milliseconds to minutes or even taken off the circuit board) compared with their high read speed (several hundred down to a few nano-seconds).

Although the two customary acronyms, ROM and PROM, are evolutionary sequential, they are not systematic. It would be more systematic if ROM is used for read-only memory and **RWM** for **read-write memory**. RWM was indeed used by a few earlier authors. The symbol conflict of W from Write and Word is avoided using the term **Program** in addition to indicating its main function as a program store. ROM and PROM will be used in this description.

For pedagogical reasons, the ROM and the PROMs including several erasable PROMs or EPROMs will be described in a sequence based on the increasing complexity of the cell structure and read-write-erase physics instead of historical evolution. Nevertheless, history and structure evolution are correlated if the history is examined [600.1] because the novel structures were evolved to give better manufacturability and higher performance (faster erase-write).

Read Only Memory (ROM)

If the storage capacitance of the Dennard DRAM cell in Fig.671.2(a) is replaced by a short circuit to represent 1 and an open circuit to represent 0, then we have a read only memory (ROM) using the MOST as the **select gate**. The open- and short-circuit bit pattern is built in or burned in during fabrication and hence cannot be changed by the user. The simplest ROM cell is just two orthogonal conductor lines, one for the bit-line (column) and the other for the word-line (row), whose cross point is either open or short circuited. The select transistor and the word line improve the sensing and random access capabilities.

Programmable Read Only Memory (PROM)

Many applications require that a bit pattern be programmed into the memory chip by the application manufacturer or the end user. This can be accomplished if a fuse is connected at each cross point between the column (bit) and row (word) lines of the ROM. The fuse can be a physically small low-value resistor that can be blown open by a current pulse or it can be a high-resistance thin insulator that can be shorted by a voltage pulse. A pointed intense laser could also be used to open or short the fuse at the cross points. This is known as the programmable read only memory (**PROM**) and more accurately, **WORM**, Write Once Read (Only) Memory. WORM is also the acronym of some first-generation rotational-disk optical memory using a very small diameter junction laser diode as the light source to write-erase as well as read. Silicon PROMs using resistor and thin-insulator fuses have been mass produced. A select MOS transistor in each cell and the word lines can again improve the sensing and random access capabilities.

Erasable Programmable Read Only Memories (EPROMs)

Nonvolatile reprogrammable silicon MOS memories, known as **EPROMs** (Erasable Programmable Read Only Memories), have been developed to allow the users to **reprogram** occasionally for updating the stored information. The three mass-produced **dielectric** Si MOS EPROMs will be described first which have reached the densities of 1Mbit and 4Mbit per chip in 1991. A promising new **ferroelectric** EPROM will also be described to finish up this section.

The information in the dielectric EPROMs is stored as electrons on a n+doped poly-Si floating gate (FG) which is insulated by surrounding SiO_2 layers to give an electron retention time greater than 10 years. A poly-Si control gate (CG) is situated on the oxide over the floating gate to select the cell for read and to select as well as to control the voltages for erase and write. This overlapping control-select gate gives the minimum geometry **one-transistor cell**. Thus, no other transistor is needed to address the cell (such as the separate select transistor in the early lower-density $\leq$1Mbit DRAM cells), albeit it is a composite and more complex transistor that performs two functions, store and address via word-select.

To program the EPROM, four **electron injection-extraction** methods via three transport mechanisms (photoemission over, injection over, and tunneling through the SiO_2/Si barrier between the FG and the Si substrate or the FG and CG;

see Table 360.1) have been used to positively or negatively charge (write) the n+poly-Si floating gate and to discharge (erase) it to the neutral state. The four methods are labeled in Figs.674.1(a) and (b) with appropriate operation biases indicated: (i) extraction by photoemission of electrons (PEE) over the SiO_2/Si barrier using ultraviolet light (UV), a very slow process (~5 minutes); (ii) avalanche electron injection (AEI) over the SiO_2/Si barrier, a medium-speed process (~10μs); (iii) channel hot electron injection (CHEI) over the SiO_2/Si barrier, a medium-speed process (~10μs); and (iv) Fowler-Nordheim tunneling electron injection (FN-TEI) or extraction (FN-TEE) through the SiO_2/Si barrier, both slow to medium-speed processes (~1s→1μs) depending on oxide thickness and electric field strength.

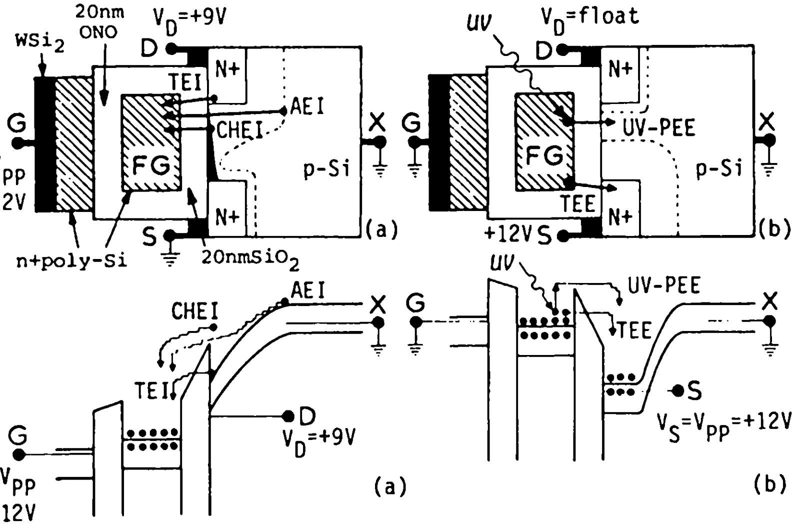


Fig.674.1 Energy band diagram and cross sectional view of the (a) three electron injection and (b) two electron extraction methods via the three electron transport mechanisms (UV-PEE, CHEI-AEI, and FN-TEE/TEI) that can be used to program (write and erase) a Si MOS EPROM cell. Mass-production choices are UV-PEE/AEI for UV-EPROM, and CHEI/FN-TEE for flash-EEPROM, whose biases are shown for write in (a) and erase in (b), and they are discussed in the text.

Hole transport mechanisms are not used due to the larger SiO_2/Si hole potential barrier (4.3eV) than electron (3.1eV) which greatly increases the hole injection-extraction times. In addition, holes generate interface traps and charged oxide traps much faster (~10 times) than electron, increasing the wearout rate and decreasing the operating life. Thus, there are six possible Si MOS NVM cell structures from the three electron injection and two electron extraction methods.

Two memory cells have been the popular choices in mass production as dictated by cost (yield) and application requirements (programming speed): the UV-EPROM (Ultra-Violet light EPROM) using CHEI to write a bit or a byte (10μs) and UV-PEE to erase the whole chip (5-minutes); and the E-EPROM, EEPROM or

E^2PROM (Electrical EPROM) using also the CHEI to write a bit or byte of cells (10μs) and FN-TEE to erase an array or a block of cells (1s), known as the **flash** EPROM (should be flash EEPROM). A third cell, the original EEPROM, uses FN for both injection (write) and extraction (erase) in order to have both high erase and write speeds (1-10μs) but it is expensive (low yield and >10 times flash EEPROM cost) due to the very-thin poly-Si tunnel oxide needed for speed.

Aside from EPROMs' intended applications as a semi-permanent store with infrequent updates, the reprogramming rate is limited by the inherently slow electron injection (write) and extraction (erase) mechanisms and also by **wearout** (**fatigue**) due to generation of interface traps by electrons and injected holes during electron injection-extraction which limits the reprogramming cycle to less than 10^6.

UV-EPROM (1971-1991!)

The first EPROM product was the Intel 2kbit UV-EPROM invented in 1971 by Frohman-Bentchkowski. The cross-sectional view is shown in Fig.674.2. It is a p-channel Si MOST with a polysilicon **floating gate** covered by an unshielded SiO_2 to facilitate UV erase. Thus, it has only three terminals (drain, source and substrate). A separate select MOST in each cell is necessary to write and read. To **program** a cell, energetic electrons, impact generated by the p-channel holes entering into the drain junction space-charge region, are injected in **avalanche** into the floating poly-Si gate ($\simeq$100μs/Byte) by the more positive FG voltage, $V_{FG} \sim -|V_D/2| \simeq -14V$, relative to the drain voltage, $-|V_D| \simeq -28V$. This produces a field-induced built-in hole channel by the negative electron charge on the floating gate. The electron trajectory is labeled by AEI (Avalanche Electron Injection) in Fig.674.2. To **erase** a cell, the electrons stored at the floating gate are released to the n-Si substrate over the SiO_2/Si barrier [$\chi(Si)-\chi(SiO_2)=4.02-0.9=3.12eV$ from Fig.412.1 and Table 413.1.] by UV light (~>4.5eV). Long exposure (5-minutes) is necessary even using the highest intensity xenon UV light source.

The present mass-produced UV-EPROM cell (1Mb introduced in 1986 and 4Mb, 1989), whose cross-sectional view looks like Fig.674.1, is a n-channel 1-**transistor** cell with a control (write) gate over the floating gate which was introduced by Intel in 1974 as a 8kbit/28V product. It was scaled down to give 256kbit chip in 1981 at a lower programming voltage (+12.5±0.5V) which has remained the current industrial standard of 1Mbit and 4Mbit UV-EPROMs. To write (or program), the **channel hot electrons** (CHE), accelerated through a 1μm channel, to the drain n+/p junction by a $V_D=+9V$, are injected over the 3.12eV SiO_2/Si barrier into the floating gate by a control gate voltage of $V_G=+12.5V$ as indicated by Fig.674.1(a). This increases the threshold voltage from +1.5V to +5V in about 100μs. The stored electron charge on the FG is sensed or read via the drain current at $V_D=+5V$ by a low gate voltage, $V_G=+2V$. Erase is by exposure of FG to UV light for 5 minutes as depicted by Fig.674.1(b). The UV light passes through the opening in the WSi_2 control gate and the quartz window in the dual-in-line package of the chip.

After twenty years of tremendous profit (1971-1991), Intel Chairman Gordon E. Moore announced on April 22, 1991 that Intel will not develop UV-EPROM chips denser than 4Mbit in favor of the flash-EEPROM because of UV-EPROM's deficiencies: tedious off-circuit-board erase, longer erase time due to smaller gate window and increasing blockage of the erase light by dust inside the chip package.

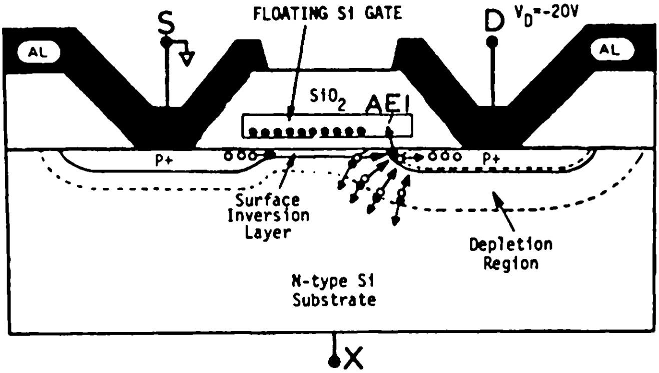


Fig.674.2 The cross-sectional view of the original p-channel Si UV-EPROM invented by Frohman-Bentchkowski in 1971. His figure in the original article (Applied Physics Letters, 18(8), 647, 15 April 1971.) is edited here to bring out at one glance the write physics due to AEI.

Flash-EPROM (a EEPROM)

The UV-EPROM is a simple and very dense (**1-transistor**) cell. But most applications require in-circuit erase for electrical reprogrammability. To achieve this, UV erase is replaced by **FN-TEE** (Fowler-Nordheim tunneling electron extraction) from the FG to the n+ source (block erase), shown in Fig.674.1(b), or n+ drain (bit erase), not shown. The current industrial standard, the 1-transistor n-channel flash EEPROM cell introduced by Intel, known as **ETOX** (**E**PROM **T**unnel **Ox**ide), whose cross-sectional view is again like Fig.674.1 and nearly identical to that of the UV-EPROM except that: (1) there is no opening in the WSi_2 CG for light, (2) the oxide between the FG and the Si substrate is reduced from 200A in UV-EPROM to 100A in ETOX flash-EEPROM for tunnel erase, and (3) the n+ source junction is diffused deeper than the n+ drain junction (not illustrated in Fig.674.1) in order to sustain the high erase voltage applied to the common source during array or block erase.

To write or program, electrons are injected onto the floating gate via CHEI just like the n-channel UV-EPROM [See Fig.674.1(a).], with $V_G=V_{PP}=+12.5V$, $V_D=+9V$, and $V_S=V_X=0V$. It takes $10\mu s$ to increase the threshold voltage from +2V to +9V. To erase, the excess electrons on the floating gate are extracted to the n+source via FN-TEE [Fig.674.1(b).] with $V_S = V_{PP} = +12.5V$, $V_G = V_X = 0$, and V_D=float. To decrease V_T from +9V to +2V takes 0.8s for $x_o = 100A$.

The programming cycle of the ETOX flash EEPROM has been increased to over 10^6 for a 1Mbit chip because good quality tunnel oxide is more easily grown on the lightly boron implanted Si substrate than on poly-Si gate and the electric field is lower due to the slower electrical erase required than in the tunnel-tunnel EEPROM. To speed up the erase, a block or an array of cells are erased simultaneously. The ETOX cell, like Fig.674.1, facilitates this operation by simultaneously biasing the common n+ source of an array or block of cells. This was first proposed by Toshiba in 1984 and coined the flash EEPROM. [IEDM Tech.Digest, 464, 1984; and ISSCC Digest of Tech. Papers, pp.168-169, 1985.] To guarantee the erase-read and program-read gate voltage margins after each programming cycle for a $V_{CC} = +5V \pm 0.5V$, an iterative on-chip algorithm was incorporated by Intel which verifies the threshold voltage of each cell by on-chip generated voltages from V_{PP} during each programming cycle. The algorithm also prevents over-erase or excessive extraction which could positively charge the FG.

The original EEPROM or E^2PROM is similar to the flash EEPROM but employs tunneling through a thinner oxide between the FG and CG or another gate for both write and erase. The very thin poly-Si oxide between the poly-Si gates has low yield, resulting in much higher cost. The speed improvement and the cost are not competitive.

F-EPROMs (FMOS and FRAM)

The third nonvolatile memory to be described employs ferroelectric polarization to store information permanently. It may become the third mass-produced nonvolatile MOST EPROM (EEPROM) because of its much higher write or program speed and its essentially permanent storage. Leading ferroelectric candidates are the Lead-Zirconium-Titanium oxide (PZT) and other perovskites. In fact, the first semiconductor memory cell ever invented was a Si MOS transistor with a ferroelectric-oxide gate proposed by I.M.Ross of Bell Labs. in 1955 [600.1]. It is shown in Fig.674.3(a) and to be denoted by FMOST (Ferroelectric MOS transistor). The second ferroelectric memory cell, invented recently (~1988), is identical to the 1-T/1-C Dennard dielectric DRAM cell except that the dielectric layer in the charge storage capacitor is replaced by a ferroelectric oxide layer and the ground plate is replaced by program (drive) lines as indicated in Fig.674.3(b) whose circuit is shown in Fig.674.3(c). This cell will be denoted by FMOSC (Ferroelectric MOS Capacitor) cell where O stands for the ferroelectric oxides. Memory chips using the FMOSC cells will be denoted by FRAM (Ferroelectric RAM). Because ferroelectric polarization can be reversed at high speed and many times without noticeable fatigue or wearout (10^{14} reversals have been reported), high speed erase and write for RAM applications should be attainable. Thus, FRAM holds great promise as a replacement of at least two hierarchies of computer memory (disk and CPU-main-memory-RAM) and perhaps even the high-speed RAM in CPU-buffer-memory. Reducing the memory hierarchy by two to three levels would revolutionize the computer architecture [600.1].

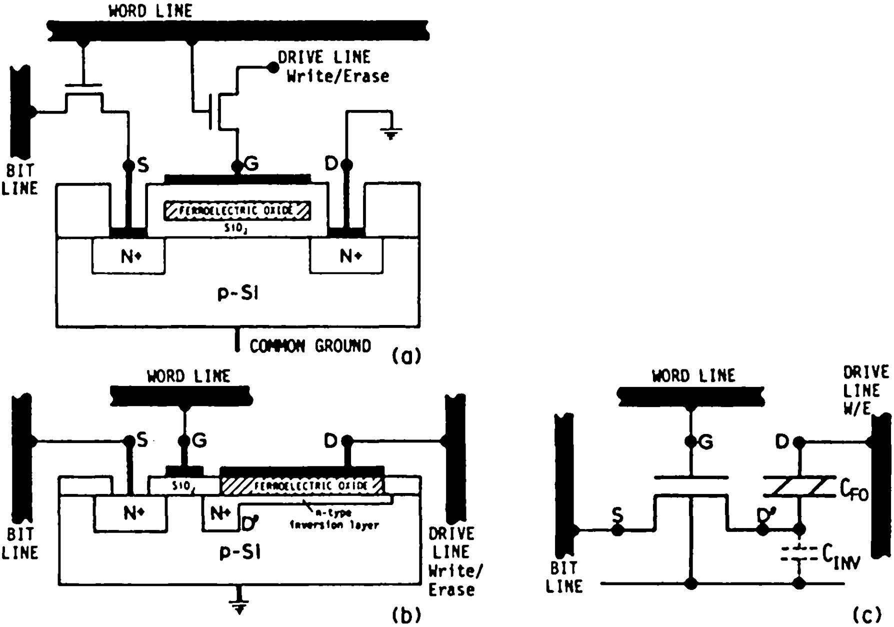

WORD LINE
DRIVE LINE
Write/Erase
BIT LINE
S
G
D
FERROELECTRIC OXIDE
SiO2
N+
N+
p-SI
COMMON GROUND
(a)
WORD LINE
S
G
D
SiO2
FERROELECTRIC OXIDE
N+
N+
D'
n-type inversion layer
p-SI
BIT LINE
DRIVE LINE Write/ Erase
(b)
WORD LINE
DRIVE LINE W/E
G
D
S
D'
CFO
CINV
BIT LINE
(c)

Fig.674.3 Cross-sectional views and circuit of the two ferroelectric SRAM cells. (a) FMOST. (b) FMOSC. (c) Circuit symbol of the metal/ferroelectric-oxide/Si capacitor memory cell.

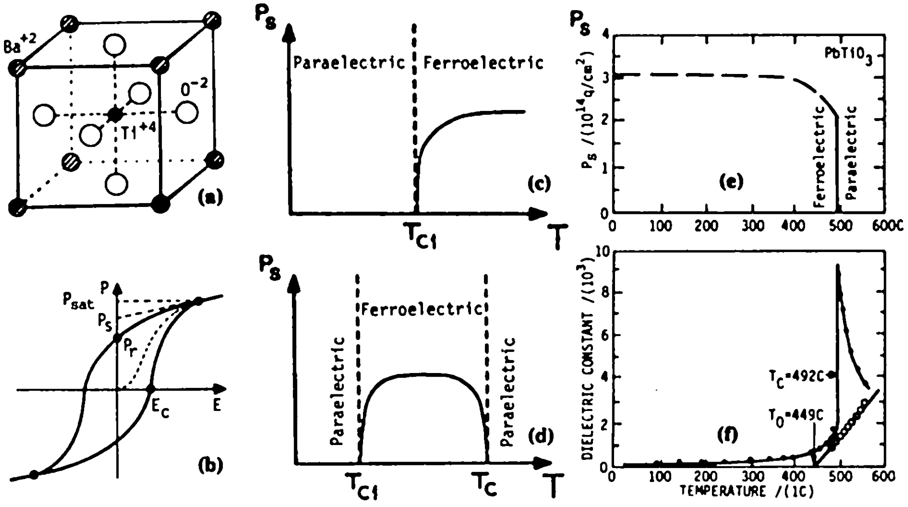
Figures 674.3(a) and (b) show that both ferroelectric SRAM MOS cells (FMOST and FMOSC) require more complex circuits to program (write/erase) than the dielectric Dennard DRAM MOS cell. As indicated in figure (b), the common capacitance plate of the Dennard DRAM cells must be replaced by isolated drive lines (DL) in order to write individual or a group of ferroelectric capacitors. The term, drive line, was chosen by the FRAM chip developers in 1990 to distinguish it from the EPROM term 'program line' because the write speed of FRAM is much faster than EPROM and the necessary write voltage may be higher. Similarly, figure (a) shows that to provide cell selectability, two series MOSTs from the word line must be used to select the gate for write/erase and the source for read.

The properties of the ferroelectric oxides have been known for decades. Current technology efforts have been focused on developing process compatibility with the Si VLSI technology. The crystal unit cell and the hysteresis loops (displacement and polarization vs electric field) of the best known ferroelectric, $BaTiO_3$, are shown in Figs.674.4(a) and (b). Temperature dependences of polarization and dielectric constant are sketched in Figs.674.4(c), (d), (e) and (f) which illustrate the **ferroelectric-paraelectric transition temperature** known as the **critical** or **Curie temperature**, T_C, and an extrapolated transition temperature, T_0, from high temperatures. $T_C > T_0$ for first order while $T_C = T_0$ for second order phase transition. Table 674.1 gives T_C, T_0 and the **spontaneous polarization** P_s of selected ferroelectric oxides. **Residual** or **remnant polarization** data is not available but is similar to the very high spontaneous polarization charge density, $P_s \approx 10^{14}$ q/cm^2. Thus, the remnant (residual) polarization (which represents the permanently stored bit) is more than adequate for permanent or static storage. The ferroelectric dielectric constant can be approximated by ε_0=constant/(T_C-T).

Table 674.1
Ferroelectric Transition Temperature and
Spontaneous Polarization of Perovskites

PEROVSKITE	CURIE T_C	EXTRAPO T_0(C)	SPONTANEOUS P_s (10^{-6}C/cm^2)	(10^{14}q/cm^2)	REMNANT P_r (10^{-6}C/cm^2)	(10^{14}q/cm^2)
**********	*****_	*******	**********__	**********	**********__	**********
$BaTiO_3$	120	109	26.0	1.6		
$SrTiO_3$	32		3	0.2		
Pb(Zr,Ti)O_3	350		6-25	0.4-1.6		
$PbTiO_3$	492	449(422)	>50	>3.1		
$LiTaO_3$	665		50	3.1		
$LiNbO_3$	1197		71(300)	4.4(19)		

( ) were older data or theory from C.Kittel, 4th ed (1971), chapter 14.



Figs.674.4 (a) The crystal structure and (b) hysteresis loops of a ferroelectric oxide, ABO_3 ($BaTiO_3$). Three possible temperature dependences of spontaneous polarization, P_s, of perovskites, (c) (d) and (e). The temperature dependence of (e) P_s and (f) dielectric constant of $PbTiO_3$.

680 BEYOND THE CONDUCTIVITY MODULATION MODEL

The ideal MOST characteristics in the preceding sections were derived using the simple conductivity modulation model which contained only the drift current and assumed $N_{AA}=N_{DD}=0$. Modification of these ideal characteristics in a real MOST will now be analyzed using the same partitioning methodology of section 641 which separated the 2-dimensional MOST into two 1-dimensional problems. Spatial inhomogeneities from $N_{AA}=N_{AA}(x,y,z)$ and $\mu_n=\mu_n(x,y,z)$ are excluded because they are not fundamental and can be solved only numerically. The first two modifications are from (i) the nonzero substrate doping ($N_{AA}\neq 0$ or $N_{DD}\neq 0$) known as the **bulk charge** or **body effects** and (ii) the **diffusion current**. They were analyzed and correlated with detailed experiments by Sah and Pao during 1964-1965. The solutions are derived in sections 681 and 682 using simple algebra to delineate and bring out the correct physics. The results are accurate for device design purposes without having to use complex numerical CAD programs which would be the next step to design today's complex and dense integrated circuits. These two effects influence the MOST characteristics so much that they must be included even in initial trial designs of MOS transistors and circuits.

The MOST characteristics are also modified by the presence of oxide and interface traps. The I_D-V_G characteristics will be parallel shifted along the V_G axis by charged oxide traps, and distorted by charged interface traps. The presence of oxide and interface traps and their density built-up with operating time are the very causes of aging of the MOSTs and failure of the MOS integrated circuits. These are discussed in section 683.

The MOST characteristics are further modified by high electric fields and applied voltages which are discussed in section 684. The high electric field effects become increasingly important when the transistor dimension decreases below one micron because the applied d.c. voltages cannot be reduced proportionally in order to retain a sufficient noise margin.

The **high electric field effects** include: (E1) drift velocity saturation due to the increasing rate of electron energy loss via optical phonon scattering in high longitudinal (or channel) electric field; (E2) electron mobility reduction due to scattering by interface defects and oxide ions when the increasingly higher transverse (gate) electric field squeezes the inversion electrons closer to the imperfect and trappy SiO_2/Si interface; (E3) increased gate (and substrate) current due to Fowler-Nordheim electron tunneling through the SiO_2/Si barrier into the oxide at increasing transverse (gate) electric field; (E4) positive threshold voltage shift due to increasing negative oxide charge from capturing tunnel-injected electrons by the oxide traps; (E5) distortion of the I_D-V_G and I_D-V_G curves due to nonuniform tunnel injection of electrons and holes into the oxide because of drain-to-source voltage drop or varying oxide thickness and dopant impurity concentration along the channel.

The high voltage effects occur in high-voltage MOSTs even when the electric field is not very high because of large dimensions. The **high voltage effects** include: (V1) channel length shortening due to thickening of the drain junction space-charge layer with increasing drain voltage; (V2) channel width widening due to fringing gate oxide electric field at high gate voltages; (V3) increasing gate (and substrate) current due to avalanche injection into the oxide over the SiO_2/Si potential barrier of interband impact generated Si electrons by high drain voltage at high channel electron current; (V4) threshold voltage shift due to: capture of the avalanche injected electrons by oxide traps, generation of oxide traps by injected and accelerated electrons that are energetic enough to break weak intrinsic and hydrogen bonds in the oxide, and generation of interface traps by the energetic injected Si electrons; and (V5) distortion of the I-V curves due to interface trap generation and spatially nonuniform trapped and generated charges in the oxide in (V4). The channel length and width effects of (V1) and (V2) are geometrical effects as well as high voltage effects and are described in the last subsection, 685.

681 Effect of Bulk Charge (Body Effects: Dopant and Substrate Bias)

Bulk charge is the term for the ionized dopant impurity in the Si surface space-charge layer underneath the gated oxide, or more generally, the impurity and majority carrier space charge in the Si surface channel next to the SiO_2/Si interface. It creates two body-effects on the current-voltage characteristics of MOSTs: it changes the threshold voltage and the shape of the I-V curves. Both were critical on the evolution of MOS integrated circuits. Both are included in today's MOS VLSI designs. The bulk charge concentration is especially crucial: it is universally used to manufacture enhancement mode nMOST on p-Si by adjusting the threshold voltage in order to cut off the built-in channel at zero gate voltage due to positive oxide charge and work function difference.

Origin of the Bulk Charge

Bulk charge was defined in (641.2C) for an nMOST on p-Si which is

$$Q_B = -\int_0^{\infty} q(P_B - P)dx < 0. \tag{641.2C}$$

It comes from the reduction and eventual depletion of the majority carriers (holes in p-Si substrate of an n-channel MOST) in the Si surface space-charge layer by the positive gate voltage and negative substrate voltage applied to an nMOST to reverse bias the surface space-charge layer. These two applied voltages move the majority carriers (holes) away from the SiO_2/p-Si interface into the quasi-neutral p-Si substrate or bulk. Simultaneously, they attract the minority carriers (electrons) to the SiO_2/Si interface to form a very thin sheet that gives the n-type surface inversion channel which produces the source-to-drain current. These voltages produce an uncompensated Si space-charge layer, $\rho \neq 0$, next to and parallel with the

SiO_2/Si interface, $x=0$. In two dimensions, the volume space charge density (Coulomb/cm^3) is given by

$$\rho(x,y,t) = q[p(x,y,t) - n(x,y) - N_{AA}] \neq 0.$$

The ρ-x space charge density and E-x energy band diagram at a y-z plane perpendicular to the channel of an nMOST are shown in Figs.681.1(c) and (d). Included are also a layer of charged interface traps whose areal density is Q_{IT} (C/cm^2), and a layer of charged oxide traps whose effective areal density at the SiO_2/Si interface is Q_{OT} (C/cm^2) whose effects are discussed in section 683.

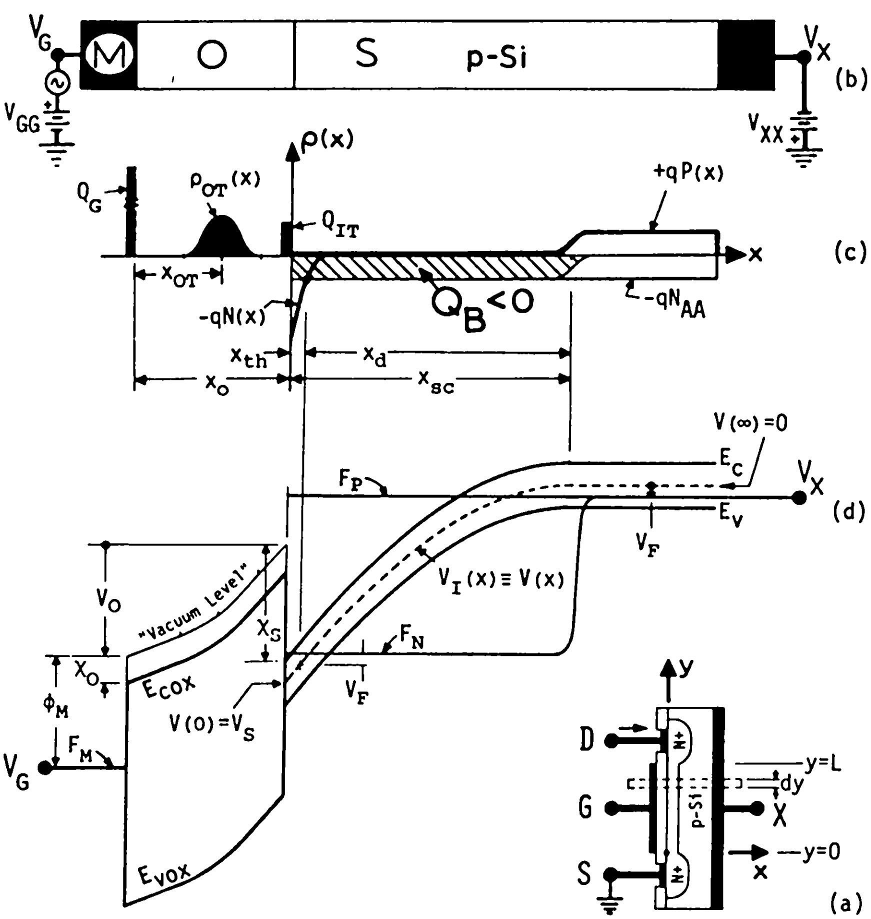

V_G
M
O
S
p-Si
V_X
(b)
V_GG
V_XX
ρ(x)
Q_G
ρ_OT(x)
Q_IT
+qP(x)
x
(c)
x_OT
-qN(x)
Q_B<0
-qN_AA
x_th
x_d
x_o
x_sc
V(∞)=0
E_C
V_X
F_P
E_V
(d)
V_F
V_I(x) ≡ V(x)
V_O
"Vacuum Level"
X_S
F_N
X_O
E_cox
V_F
φ_M
V(0)=V_S
F_M
V_G
E_vox
y
D
n+
y=L
dy
G
p-Si
X
S
n+
y=0
x
(a)

Fig.681.1 The nMOST on p-Si. (a) Cross-sectional view. (b) Expanded view of the slice of y-z plane cut out of the channel (dashed box in (a). (c) Charge distribution and their labels. (d) The E-x energy band diagram of a channel cross section at a distance y from the source.

The p-Si wafer of an nMOST biased into the active region (i.e. with channel current) can be partitioned into three x-layers, consisting of the two sub-layers of the Si surface **space-charge layer** (SCL) and the quasi-neutral bulk layer. The two sub-layers of the p-Si surface SCL are: (i) the **strong inversion layer** ($N>n_{th}=P_B\simeq N_{AA}>P$ in $0+<x<x_{th}$) at the SiO_2/Si interface, and (ii) the **depletion** layer, x_d, which ends at the edge of the **surface space-charge layer** whose thickness is labeled x_{sc} ($=x_{th}+x_d$) and defined by $P<<P_B\simeq N_{AA}$ in $0<x<x_{sc}$. Note majority carriers (holes) are depleted in the entire space-charge layer, $0+<x<x_{sc}$, which includes the strong inversion layer. However, minority carriers (electrons) accumulate in a thin sheet at the SiO_2/Si interface whose concentration increases from a very low bulk value at ($x=\infty,y=0$), $N_B\simeq n_i^2/N_{AA}$, to a high value, $N(0,0)$, at the SiO_2/Si interface near the source junction. Along the channel at the SiO_2/Si interface, $N(x=0,y)$ decreases along the channel (y-direction) towards the drain, $N(0,y)<N(0,0)$, because of the reverse bias applied to the drain n+/p junction. $N(0,L)\to 0$ at the edge of the drain n+/p junction space-charge layer and $N(x,L)=0$ when $V_D\geq V_G-V_{GT}$. (iii) The rest of the Si thickness is the **quasi-neutral** p-type **bulk layer** with $P=P_B\simeq N_{AA}>n_i>N$ and enclosed by $x_{sc}<x<T_p-x_{sc}$ where T_p is the thickness of the p-Si wafer. The three layers are clearly shown in Figs.681.1(c) and (d). Simplification of these diagrams was made in Fig.641.2(b) and used to derive the simple one-dimensional MOS differential equation, (643.1).

The bulk charge defined by the just cited integral, (641.2C), is the areal charge density (C/cm^2) of either an excess or deficit of majority carrier (hole) space charge relative to the flat-band value in the surface space-charge layer, $0+<x<x_{sc}$. The integral is the shaded area given in Fig.681.1(c) labeled Q_B which is negative because holes are depleted in the surface SCL $0\leq x\leq x_{sc}$.

Analytical Approximation of Bulk Charge (Depletion Model)

An analytical approximation to the Q_B integral, (641.2C), can be obtained using the depletion approximation since Q_B comes mainly from the depletion layer of the surface SCL as indicated by Fig.681.1(C). Thus, by inspection,

$$Q_B = -\int_0^\infty q(P_B - P)dx \tag{641.2C}$$

$$\simeq -\int_0^{x_{sc}} q(P_B - 0)dx = -qP_B\cdot x_{sc} \simeq -qN_{AA}\cdot x_{sc}. \tag{681.1}$$

The strong inversion layer is very thin, $x_{th}<<x_{sc}$, thus, $x_{sc}=x_d+x_{th}\simeq x_d$. x_d can be approximated by that of a reverse-biased n/p junction diode, (531.1), since it was obtained by integrating the Poisson equation in the SCL assuming with $N_{DD}\neq N_{DD}(x)$, $N_{AA}\neq N_{AA}(x)$ and $N=P=0$. For MOST, x_{sc} is located entirely in the semiconductor (p-Si), thus,

$$x_{sc}(y) \approx x_d = x_{pn} = \sqrt{2\varepsilon_s[V_{bi}+V_R]/qN_H} \quad \text{(depletion approximation)} \quad (531.1)$$
$$= \sqrt{2\varepsilon_s[V_B(y)]/qN_H} \quad \text{(depletion approximation)} \quad (681.1A)$$
$$= \sqrt{2\varepsilon_s[V_{IO}-V_X+V_N(y)]/qN_{AA}} \quad \text{(constant QFL in x)} \quad (681.1B)$$
$$= \sqrt{2\varepsilon_s[2V_F-V_X+V_N(y)]/qN_{AA}}. \quad \text{(strong inversion)} \quad (681.1C)$$

The total Si energy band bending or **potential barrier height** of the surface space-charge layer, $V_B(y)$, (known also as the **surface potential** in MOSC theory given in Chapter 4) consists of the built-in barrier height, $V_{bi}=V_{IO}$, and the reverse applied voltage, $V_R=V_N-V_P$, across the barrier or the space-charge layer. These give (681.1B). This is precisely the relationship used in Chapter 5 to analyze n/p junction diode characteristics and known as the **constant quasi-Fermi level (QFL) approximation**, i.e., constant in x only in the 1-d n/p junction space-charge layer. As indicated in Fig.681.1(d), the majority carrier (holes) quasi-Fermi level, $F_P=-qV_P$, is spatially constant: $V_P(x,y)=\text{constant}=V_X=-V_{XX}$(applied substrate bias). In contrast, the minority carrier (electrons) quasi-Fermi level, $F_N=-qV_N$, is independent of the thickness of the surface space-charge layer: $V_N(x,y)\neq f(x)$, but it does change along the channel, from $V_N(y=0)=V_S$ at the source junction to $V_N(y=L)=V_D$ at the drain junction. Thus, $V_N(x,y)=V_N(y)$ is **not constant in y** in the 2-d space-charge layer of an nMOST. In the simple analyses given in sections 64n, we used the symbol $V(y)$ for $V_N(y)$.

The surface potential at the source in (681.1B) is approximated by $V_{IO} \approx 2V_F$ to give (681.1C) based on the **strong inversion model** which was used in chapter 4 to compute the high-frequency capacitance in strong inversion. This $V_{IO}\approx 2V_F$ approximation has a simple physical significance in MOST. It is the Si surface potential barrier height at the source that gives $N(x=0,y=0) = P_B$ (or $\approx N_{AA}$), i.e., the minority carrier (electron) concentration at the source end of the surface channel is equal to the majority carrier (hole) concentration in the bulk. Thus, in the strong inversion model ($V_{IO}=2V_F$), the analytical expression for the bulk charge in the depletion approximation is

$$Q_B(y) \approx -\, qN_{AA}x_{sc} \quad (681.2)$$
$$= -\sqrt{2\varepsilon_s qN_{AA}[V_B]} \quad \text{(depletion approximation)} \quad (681.2A)$$
$$= -\sqrt{2\varepsilon_s qN_{AA}[V_{IO}-V_X+V(y)]} \quad \text{(constant QFL in x)} \quad (681.2B)$$
$$\approx -\sqrt{2\varepsilon_s qN_{AA}[2V_F-V_X+V(y)]}. \quad \text{(strong inversion model)} \quad (681.2C)$$

The three terminal voltages, $V(y=0)=V_S$, $V(y=L)=V_D$, and substrate bias V_X, are all measured with respect to a common but arbitrary reference voltage node. In Figs.681.1(a)-(d), we have let $V_S=0$ and $V_X=-V_{XX}<0$. In the following analysis, we let $V_X=0$ and $V_S>0$.

Body Effects on Threshold Voltage

The gate threshold voltage to cut off the channel conductivity at the source junction is given by

$$V_{GT} = V_{IO} + V_S + V_{GTX} + V_{FB} \quad \text{(general)} \tag{681.3}$$

$$= 2V_P + V_S + V_{GTX} + V_{FB} \quad \text{(strong inversion model).} \tag{681.3A}$$

Referring to the source terminal and using the symbol V_{GTS} or V_{GST}, then

$$V_{GTS} \equiv V_{GT} - V_S = V_{IO} + V_{GTX} + V_{FB} \equiv V_{GST} \quad \text{(general)} \tag{681.4}$$

$$= 2V_P + V_{GTX} + V_{FB} \quad \text{(strong inversion).} \tag{681.4A}$$

The two contributions are as follows. V_{FB} is the **flat-band voltage** defined by

$$V_{FB} = \phi_{MS} - (Q_{IT} + Q_{OT})/C_o \tag{681.5}$$

which was also identically defined for a MOSC by (410.32). V_{GTX} comes from the bulk charge and is defined by (V_S-$V_X \geq 0$ for reverse or zero bias on the S/X n+/p junction)

$$V_{GTX} = - Q_B/C_o \approx + qN_{AA} \cdot x_{SC}/C_o \tag{681.6}$$

$$= + (1/C_o)\sqrt{2\varepsilon_s qN_{AA}[V_B]} \quad \text{(depletion approx.)} \tag{681.6A}$$

$$= + (1/C_o)\sqrt{2\varepsilon_s qN_{AA}[V_{IO}+V_S-V_X]} \quad \text{(constant QFL } V_P \text{ and } V_N \text{ in x)} \tag{681.6B}$$

$$\approx + (1/C_o)\sqrt{2\varepsilon_s qN_{AA}[2V_P+V_S-V_X]} \quad \text{(strong inversion).} \tag{681.6C}$$

Three results are noted. (i) The body effect increases the threshold voltage. The **negatively charged** acceptor atoms in the surface space-charge layer gives a **positive shift** of the threshold voltage since additional gate voltage must be applied to overcome the negative charge before the gate voltage can attract electrons to the surface to form the conduction channel. This was given by $V_{GTX} = +|-Q_B/C_o|$. In production of nMOSTs on p-Si, this is used to compensate for a **negative shift** due to (a) **positive oxide charge**, $-Q_{OT}/C_o$, and (b) metal/Si work function difference, ϕ_{MS}. In pMOST on n-Si, the threshold-voltage-shift due to bulk charge is negative because the donor impurities are positively charged. This negative shift can be used to compensate for the positive shift from negative oxide charge and work function difference.

(ii) V_{GT} obtained in (681.3) is the gate threshold voltage measured relative to the source as indicated by (681.4). The body contribution from (681.6C) depends only on the voltage difference between the source and the substrate, V_S-V_X, i.e. the reverse bias applied to the source/substrate n+/p junction. This is obvious from (681.6B) and (681.6C) which can be rewritten as

$$V_{GTX} = -\ Q_B/C_o$$

$$= +\ (1/C_o)\sqrt{2\varepsilon_s q N_{AA}(V_{IO}+V_S-V_X)} \quad \text{(depletion approx)} \qquad (681.6B)$$

$$\equiv V_{GTXS} \equiv V_{GTSX} \equiv V_{GSTX} \quad \textbf{(unified multiple notation)} \qquad (681.7A)$$

$$V_{GTSX} \equiv +\ (1/C_o)\sqrt{2\varepsilon_s q N_{AA}(V_{IO}+V_{SX})} \quad \text{(depletion approx)} \qquad (681.7B)$$

$$\simeq +\ (1/C_o)\sqrt{2\varepsilon_s q N_{AA}(2V_F+V_{SX})} \quad \text{(strong inversion).} \qquad (681.7C)$$

The threshold voltage given in the form of (681.4) or

$$V_{GTS} = V_{IO} + V_{GTSX} + V_{FB} \equiv V_{GST} \qquad (681.8)$$

is particularly useful when the source and drain junctions are dissimilar due to fabrication differences (x_o, N_{AA}, ϕ_{MS}, Q_{OT}, Q_{IT}) or high electric field stress during operation (different Q_{OT} and Q_{IT} build-up with operating time). For any asymmetrical source and drain, this threshold formulae can be immediately used to generalize the MOST device current equation into a symbolic-symmetrical form because the source and drain terminal labels are arbitrary and interchangeable. This property was used to write down the symmetrical MOST current equation given by (663.1), termed the **device equation**. The symmetry has tremendously simplified the physical reasoning required for simultaneous solution of the highly nonlinear device equation with the **linear circuit equations** without having to use one of the terminals, G, D, and S, as the reference node. Section 663 gave examples of using this symmetry to quickly obtain the transient voltage and current waveforms and the loci on the nonlinear device i_D-v_D and i_D-v_G planes of charging and discharging capacitors through a MOST switch.

(iii) The strong inversion approximation, (681.6C) and (681.7C) using $V_{IO} \simeq 2V_F$, although adequate to describe the body effect above the threshold voltage, is inadequate to characterize the body effects in the subthreshold current-voltage range. This is because the subthreshold current is dominated by diffusion or the number of electrons that climbed over the potential barrier of the n+/p source junction into the surface channel. It is proportional to the electron density at the source end of the channel, $I_D \propto Q_N(y=0) \propto \exp(qV_{IO}/kT)$. Thus, the dependences of I_D on V_G, V_S and V_X would be missed if V_{IO} is taken as a constant, $2V_F$. In the next section, 682, a better approximation of V_{IO} is derived in the subthreshold range defined by $0 < V_{IO} < 2V_F$. That derivation starts with (681.6B) and (681.7B) without making the strong inversion approximation, $V_{IO} \simeq 2V_F$.

Body Effects on I-V Shape

The position or channel potential dependence of the bulk charge $Q_B(y)$ will give an additional voltage dependent drain current term when the MOST differential equation, (643.1), is integrated along the channel from the source, $y=0$, to the drain, y=L. The simplified MOST equation, (643.3), assumed $Q_B(y)$=constant and lumped it into V_{GT} defined by (642.3). Removing this assumption by writing out the bulk charge contribution explicitly, (643.1) becomes

$$\begin{aligned} I_D &= \mu_n Q_N (dV/dy) Z \\ &= \mu_n C_o [V_G - V_{IO} - V + (Q_B/C_o) - V_{FB}](dV/dy)Z \qquad (681.9) \end{aligned}$$

where $V_{IO}=2V_F$ in the strong inversion range. The MOST current-voltage equation (643.2) was obtained by integrating (681.9) over the length of the channel when the bulk charge is assumed constant. When the bulk charge varies along the length of the channel, (681.9) gives an additional voltage-dependent drain current term from the bulk charge which is

$$\begin{aligned} I_D(Q_B) &= \mu_n (Z/L) \int_{V_S}^{V_D} Q_B dV \\ &= -\ \mu_n (2Z/3L) \sqrt{2\varepsilon_s q N_{AA}} \left[(V_{IO}+V_D-V_X)^{3/2} - (V_{IO}+V_S-V_X)^{3/2} \right]. \qquad (681.10) \end{aligned}$$

When $V_X=V_S=0$, this simplifies to

$$I_D(Q_B) = -\ \mu_n (2Z/3L) \sqrt{2\varepsilon_s q N_{AA}} \left[(V_{IO}+V_D)^{3/2} - (V_{IO})^{3/2} \right]. \qquad (681.10A)$$

Added to (643.2), (650.1), (663.1) or (670.2) the total drain current is

$$I_D = (Z/2L)\mu_n C_o [(V_G-V_T-V_S)^2 - (V_G-V_T-V_D)^2] + I_D(Q_B). \qquad (681.10B)$$

where $V_T=V_{IO}-V_{FB}$ as indicated in (681.9) and $V_{IO}=2V_F$ in the strong inversion range. The above result shows that the negative bulk charge from negatively charged acceptor impurity ions in the surface SCL of the p-Si reduces the drain current at a given V_G. This is expected and consistent with the increasing threshold voltage with higher bulk charge since an additional positive gate voltage is needed to provide the additional electric flux to terminate at the negative acceptors ions before V_G can induce electron charge in the channel. Note that the bulk charge also causes the drain saturation to occur sooner because the reverse bias on the drain junction, V_D-V_X, thickens the space-charge layer of the channel near the drain junction and causes the saturation condition, $Q_N(V_G,V_D)=0$, to occur at a lower V_D than when bulk charge is assumed constant.

Five Illustrations and a Numerical Example of the Body Effect

The effects of the bulk charge on the transfer characteristics, I_D versus V_G, are illustrated in Fig.681.2 by five nMOST examples. The curves are in the drain current saturation range, $V_D>V_G-V_{GT}$, where the drain current varies with V_G and not V_D. The straight line part of curve 0 is in the nonsaturation range, V_G-$V_{GT}>V_D=5$V. Curve 0 is an nMOST with no bulk charge ($N_{AA}=N_{DD}=0$) and no metal/semiconductor work function difference ($\phi_{MS}=0$). Its I_D-V_G curve is given by $I_D/(\mu_n C_o Z/L)=V_G^2/2$. Note that the straight line in the non-saturation range intercepts the V_G axis at $V_G=V_D/2=2.5$V. The effect of a work function difference is given by Curve 1 which parallel shifted from curve 0 by $\phi_{MS}=1$V. Curve 2 shows an additional parallel positive gate voltage shift of $+|-Q_{OT}|/C_o$ when a negative oxide charge of Q_{OT}(C/cm^2) is present. Curve 3 shows the effect of bulk charge due to ionized and negatively charged acceptors in the surface space-charge layer of the p-Si substrate with $N_{AA}=10^{16}$cm^{-3} and $x_o=1000$A. Curve 4 shows the effect of a large reverse substrate bias voltage of V_X = -25V_P $\approx$ -10V.

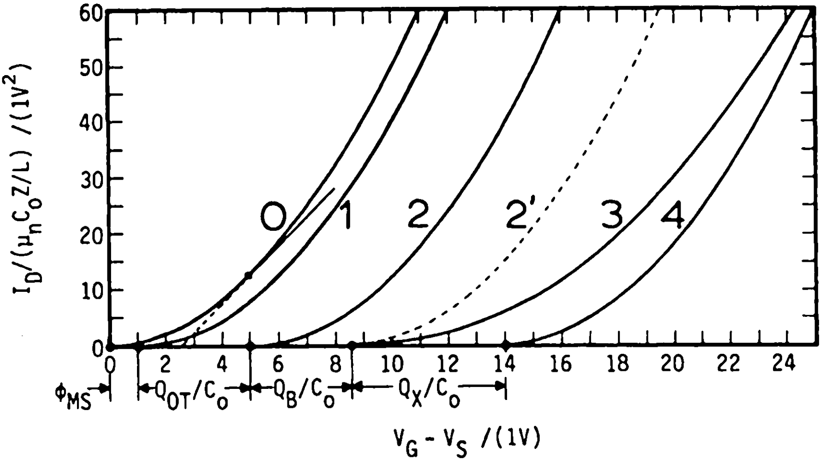


Fig.681.2 The transfer characteristics, I_D vs V_G, of nMOSTs showing the effects of bulk charge on the characteristics near the threshold voltage. Curves 0 to 4 are discussed in the text.

As stated at the beginning of this section, there are actually two effects from the bulk charge which makes curve 3 different from curve 2. For ease of comparison, curve 2 is shifted to 2'. The first effect is the positive shift of the threshold voltage which is given by (681.6C). The second effect is a slower and non-parabolic rise of the drain current with gate voltage because part of the gate voltage or charge is used to push the holes away into the p-Si bulk. The current reduction was given by (681.10A).

At large reverse substrate bias, $V_X<<0$, the drain saturation current again increases parabolically with gate voltage, as shown by curve 4 in Fig.681.2. In the asymptotic limit of very large reverse substrate bias, the parabolic dependence become $I_D/(\mu_n C_o Z/L)=\frac{1}{2}(V_G-V_{GTS})2$ where threshold voltage, V_{GTS}, is given by (681.8). This is illustrated by curve 4.

To illustrate the body effect further, a numerical example is now given using values given in sections 644 and 653 with $N_{AA}=10^{16}\text{cm}^{-3}$. The thickness of the surface space-charge layer at zero applied voltage can be calculated from (681.1C). The surface potential or potential drop across the space-charge layer is

$$V_{IO} = 2V_F = 2(kT/q)\log_e(N_{AA}/n_i) \tag{681.11}$$

$$= 2\text{x}0.02585\text{x}\log_e(10^{16}/10^{10}) = 0.714\text{V}, \tag{681.11A}$$

so the space-charge layer thickness at $V_S=V_X=0$ is

$$x_{sc} = \sqrt{2\varepsilon_s V_{IO}/qN_{AA}} \tag{681.12}$$

$$= \sqrt{2\text{x}11.7\text{x}8.854\text{x}10^{-14}\text{x}0.714/1.6\text{x}10^{-19}\text{x}10^{16}} = 3040\text{A}. \tag{681.12A}$$

The bulk charge density at zero bias is then

$$Q_B = qN_{AA}x_{sc} = 3.04\text{x}10^{11}q = 4.87\text{x}10^{-8}\ \text{Coulomb}. \tag{681.13}$$

Assuming $x_o=100\text{A}$ for the state-of-the-art one-micron nMOST, then

$$C_o = \varepsilon_o/x_o = 3.9\text{x}8.85\text{x}10^{-14}/10^{-6} = 3.45\text{x}10^{-7}\ \text{F/cm}^2. \tag{681.14}$$

This gives a bulk charge contribution to the threshold voltage of

$$V_{GTX}(Q_B,V_S=V_X) = Q_B/C_o = 4.87\text{x}10^{-8}/3.45\text{x}10^{-7} = 0.141\text{V}. \tag{681.15}$$

The voltage magnitude seems small, but it is important on the performance of high density chips which contain many transistors. Its significance can be further recognized by writing the total bulk charge in units of electron charge, q. Then (681.13) gives $3.0\text{x}10^{11}$ cm^{-2} which is 10^{-3} monolayer of impurity on a Si surface with $\sim 4\text{x}10^{14}\text{Si/cm}^2$ as estimated in chapter 1. It is this small dose value that dictates the development of highly accurate ion implantation technology to enable production control of the threshold voltage, via bulk charge doping, of both the nMOST and the pMOST in the CMOS. Monolithic CMOS Si integrated circuits could not be manufactured as a competitive product until 1970 when production ion implanter became available. The first inexpensive product is the digital wristwatch that used low power CMOS for frequency division down to 1Hz from several hundred kHz to a few MHz generated by a CMOS quartz oscillator [600.1].

682 Subthreshold Characteristics (Diffusion Current)

The drain current of the nMOST is not cut off sharply at the strong inversion threshold gate voltage, V_{GTS}, which was defined in (681.4A) by assuming V_{IO} in (681.4) is a constant and equal to $2V_F$

$$V_{GTS} = V_{GT} - V_S = V_{IO} + V_{FB} + V_{GTSX} \qquad \text{(general)} \tag{681.4}$$

$$\approx 2V_F + V_{FB} + V_{GTSX}. \qquad \text{(strong inversion)} \tag{681.4A}$$

There is still a minute electron channel current flowing from the source to drain when the applied gate voltage is below the strong-inversion threshold gate voltage, $V_{GS}<V_{GTS}$, or $V_{IO}<2V_F$. This is known as the **subthreshold current**. It arises from minority carriers (electrons in nMOST on p-Si), injected by the source n+/p junction, diffusing through the surface channel, and collected by the reverse biased drain n+/p junction. It is precisely the very same minority carrier injection-diffusion-collection mechanism that makes a bipolar junction transistor (BJT) work which will be studied in the following chapter. A **geometrical difference** is that in the subthreshold nMOST, the p-type base region is a thin and fairly long surface space-charge channel layer which may become intrinsic and even slightly n-type; while in the BJT, the p-type base is a wide and fairly short (or thin) bulk quasi-neutral acceptor-doped p-Si bulk layer. **The major difference, however, lies in the control of minority carrier injection by lowering the barrier height of the n+/p/n BJT emitter and nMOST source n+/p junction.** In BJT, it is lowered by the forward bias voltage applied to the emitter/base n+/p junction. In nMOST, it is lowered near the SiO_2/Si interface at the source junction by the gate voltage or transverse electric field in the gate oxide while the n+/p source/substrate junction itself could even be reverse biased. Obviously, the nMOST can operate as an n+/p/n+ BJT if the n+/p source junction of the nMOST is forward or zero biased, $V_S-V_X\lesssim 0$. Indeed forward-biasing the n+/p source junction has been used in VLSI circuits to make the nMOST operate as a **lateral BJT**.

To illustrate the subthreshold current, we first plot the I_D-V_G curves of an nMOST on linear scale as shown in Fig.682.1(a) in saturation, $V_D-V_S\gtrsim V_G-V_S-V_{GTS}$ where $V_{GTS}=V_{TH}=1.0V$. The numerical material and device parameters are given in the figure caption. The drain current seems to cease when $V_{GS}=V_{GTS}$. However, when plotted on a semilogarithmic scale as in Fig.682.1(b), the drain current is not cut off at V_{TH} but drops exponentially with a nearly constant slope. For the curve of the thinner oxide (10nm) and large substrate bias, $V_S-V_X=10V$, the **subthreshold slope** approaches $(kT/q)\log_e 10 = 59.52\simeq 60$mV/decade. For thicker oxide (100nm) and smaller substrate bias, $V_S-V_X=0$ to 5V, the subthreshold cutoff is not as sharp, almost 120mV/decade at $V_S-V_x=0$.

In order to simplify the algebra and focus on the physics, the subthreshold current was neglected in the preceding analyses on the bulk-charge effects. This was justifiable because when the surface channel is strongly inverted to n-type, the electron drift current is so much higher than the electron diffusion current through the n+/n/n+ channel. The channel current was defined to begin abruptly at $V_{IO}=2V_F$ by $I_D \propto Q_N=q\int(N-N_B)dx$, which is the reason why the strong inversion curve in Fig.682.1(b) drops to zero asymptotically at $V_{GS}=V_{GST}\equiv V_{GTS}$. Thus, a threshold voltage was tacitly defined when $V_{IO}=2V_F$ is assumed in the threshold voltage formulae (681.4) to give (681.4A). $V_{IO}=2V_F$ was also used to compute the surface barrier heights: $V_B=V_{IO}+V_S-V_X$ at the source, and $V_B=V_{IO}+V_D-V_X$ at the drain; the surface barrier thickness in (681.1C), $x_{sc}=\sqrt{2\varepsilon_s V_B/qN_{AA}}$; the bulk charge $Q_B=-qN_{AA}x_{sc}$ in (681.2C); and the electron charge density, Q_N in (681.9).

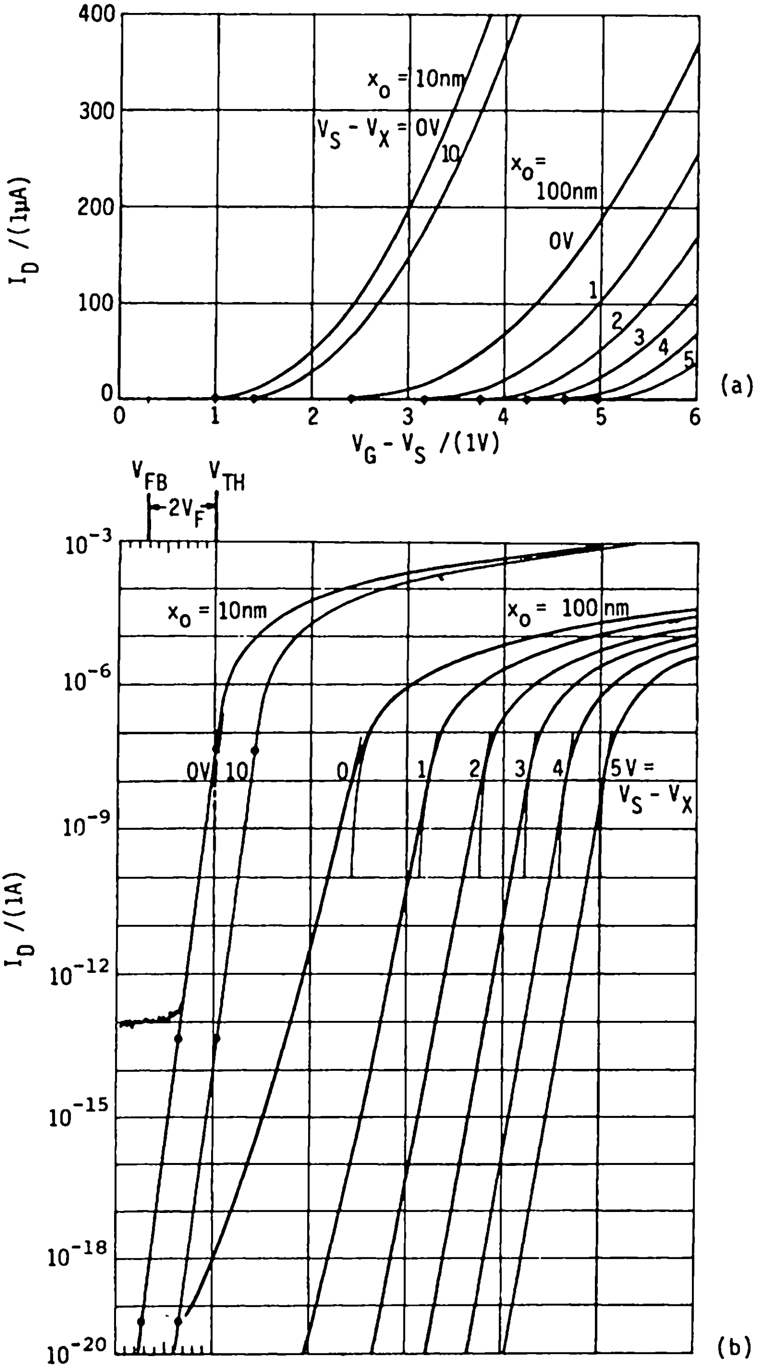


Fig.682.1 Subthreshold transfer characteristics of an Si nMOST, I_G vs V_G, for $V_D > 4kT/q$ ≈100mV. Inputs are: $Z/L = 1$, $N_{AA} = 10^{16}\text{cm}^{-3}$, $x_o = 100\text{A}$ and 1000A, $\varepsilon_o = 3.9$, $\varepsilon_s = 11.7$, $T = 300\text{K}$, $n_i = 10^{10}\text{cm}^{-3}$, $\mu_n = 289.6 \approx 300\text{cm}^2/\text{V-s}$, $2V_F = 0.7142\text{V}$, $V_{FB} = 0.1448\text{V}$, and $V_{GTX}(V_X = 0) = 0.1410\text{V}$ to make $V_{FB} + 2V_F + V_{GTX}(V_X = 0) = 1.0\text{V}$. (a) Linear. (b) Semilog.

Although the channel and drain current is low in the subthreshold range, circuit applications demand that its dependence on the gate voltage in the subthreshold range be known to assure complete cutoff of the channel, source and drain currents. Otherwise, there may be a significant standby power dissipation, signal leakage or cross talk, and noise. Attempts have also been made to design MOSTs to operate in the subthreshold range for low power applications which would require an accurate design equation of subthreshold I_D versus V_G.

Definition of the Subthreshold Range

In the subthreshold range, the Si surface channel is barely formed at the SiO_2/Si interface. The channel may not even be inverted, i.e. still p-type or $P(x=0,y)>n_i>N(x=0,y)$, giving a bipolar-transistor-like n+/p/n+ surface path. The subthreshold range can be quantitatively defined by $0 \lesssim V_{I0} \lesssim 2V_F$ or the corresponding gate voltage range given by $(V_{FB}+V_{GTX}) \lesssim V_{GS} \lesssim V_{GST}=(V_{FB}+V_{GTX})+2V_F$. Thus, the **flat band voltage** defined in (410.32) for MOSC is precisely also the **subthreshold voltage** of the MOST:

$$V_{THsub} = V_{FB} = \phi_{MS} - (Q_{OT} + Q_{IT})/C_o. \qquad (681.5)$$

It is not a coincidence to have the flat-band as the lower boundary of the subthreshold range because it is the gate voltage of true cutoff of the channel current since the channel conduction carrier (electron) density is defined by $Q_N = \int q(N-N_B)dx$ and the channel current is given by $I_D = \mu_n Q_N (dV/dy)$. Thus at flat band, $N=N_B$, so $Q_N=0$ and $I_D=0$. This definition does not predict the minute diffusion current in the majority carrier accumulation range when $V_{I0}<0$ and $N<N_B$, but this current is extremely small and usually masked by leakage and noise as shown in Fig.682.1(b).

Intrinsic Surface - Onset of Drift Current

Another characteristic gate voltage in the subthreshold range, useful here and also used in chapter 4 for MOSC's, is the **intrinsic gate voltage**, V_{GI}, defined by (411.34). It corresponds to a surface potential of $V_{I0}=V_F$ which is the **intrinsic surface** condition at the SiO_2/Si interface $(x=0)$: $N(x=0)=n_i=P(x=0)$. It is the gate voltage at which inversion starts. The inversion begins at the SiO_2/Si interface boundary $(x=0)$ of the surface channel when $V_{I0}=V_F$ and $V_{GS}=V_{GI}$. It spreads into the Si bulk when V_{GS} increases above V_{GI} and V_{I0} increases above V_F. Above V_{GI} and V_F, the drift component of the channel current begins and increases rapidly with increasing V_{GS} since the n+ source and drain are now connected by an increasingly conductive thin n-type surface inversion layer which gives the n+/n/n+ surface conduction path. As V_{GS} increases, the source junction of the surface channel changes from n+/p to n+/i, to n+/n-, and finally to n+/n when the strong inversion condition $V_{I0}=2V_F$ is reached. The intrinsic condition is the midpoint between the flat band or drain current cutoff $(V_{GTX}+V_{FB})$ and the strong inversion point $V_{GTS} = (V_{GTX}+V_{FB})+2V_F$. It is labeled on the two 100nm-oxide subthreshold lines ($V_{SX}=0$ and 10V) .in Fig.682.1(b) which show that the drain current below the intrinsic gate voltage is masked by noise and leakage current.

Analysis of the Subthreshold Current

Three mechanisms control the source-to-drain channel current, each in one of the three layers, and each will give the source or drain terminal current if solved. However, the current can be most easily calculated from the diffusion current flowing in the channel which assumes that the drift current can be neglected. Thus,

$$J_{Ny} = qD_n(dN/dy) + q\mu_n N(-dV/dx) \approx qD_n(dN/dy).$$

Note that the drift current was the sole component used to calculate the channel and drain current in the strong inversion range when $V_{GS} > V_{GTS}$ and $V_{IO} = 2V_F$. Using the diffusion model, the subthreshold current due to electron diffusion through the channel is given by

$$I_D = -\int_0^{\infty} J_{Ny}dxZ \approx -\int_0^{\infty} qD_n(dN/dy)dxZ \tag{682.1}$$

$$\approx -\int_0^{\infty} qD_n(d/dy)(N-N_B)dxZ \approx -ZD_n(d/dy)\cdot\int_0^{\infty} q(N-N_B)dx \tag{682.1A}$$

$$= -ZD_n(d/dy)\cdot Q_N \tag{682.1B}$$

which can be integrated immediately to give

$$I_D = (Z/L)D_n[Q_N(y=0) - Q_N(y=L)] \tag{682.2}$$

where the electron charge density was defined by

$$Q_N(y) = q\int_0^{\infty} [N(x,y) - N_B]dx. \tag{682.3}$$

Note that the subthreshold diffusion current given above is proportional to Q_N while the superthreshold drift current given by (681.9) and (643.2A) is proportional to the integral of Q_N over the channel voltage drop, $\int Q_N(V)dV$. It is evident that the MOST differential equation can be set up to include both components. This was solved by Pao and Sah in 1964 by double numerical integration. In the present analyses, it is given by the sum of the drift current equation, (681.9), and the diffusion current equation to be obtained in the following analysis.

To get an analytical formulae for the subthreshold current from (682.2) we need an analytical formulae of the electron density, Q_N. We give a detailed description in the following paragraph on how to derive this analytical formulae since it has been elusive to many textbook authors and researchers.

We already used an analytical formulae for Q_N in the strong inversion range which was given by (681.9):

$$Q_N = C_o(V_G - V_B - V_X + Q_B/C_o - V_{FB}) \tag{682.4}$$

where

$$V_B(y) = V_{IO} + V(y) - V_X \tag{682.4A}$$

$$= 2V_F + V(y) - V_X \quad \text{(strong inversion)} \tag{682.4B}$$

is the barrier height across the channel thickness in the x-direction. It is the sum of two terms at position y: the barrier height, V_{IO}, when $V(y)-V_X=0$, and the reverse bias across the channel thickness, $V(y)-V_X$. Equation (682.4) is obtained by going through the E-x energy band diagram at a position y in the channel as shown in Fig.681.1(b). It is strictly correct in our two 1-d models for the superthreshold range where we approximated V_{IO} by $2V_F$. However, it cannot be used to get an analytical formulae of $Q_N(V_S)$ and $Q_N(V_D)$ in the subthreshold range because Q_N is so much smaller than the other terms in (682.4), such as C_oV_G, so that it is essentially zero as a first approximation. Furthermore, although Q_N is very small, it varies exponentially with V_{IO}. Thus, the assumption of $V_{IO}=2V_F$ can only be made in the superthreshold range. It is this very change of V_{IO} with V_S, V_X and V_G over the subthreshold range $0 \le V_{IO} \le 2V_F$, that gives the voltage dependences of the subthreshold drain current. Thus, we must find a second equation of Q_N as a function of V_{IO} so that it can be solved simultaneously with (682.4) to give V_{IO} and Q_N as a function of V_G, V_S and V_X.

As indicated below, Q_N is dropped in (682.6) to give (682.6A) because it is very small compared with C_oV_G in the subthreshold range.

$$Q_N = C_o(V_G - V_B - V_X + Q_B/C_o - V_{FB}) \tag{682.5}$$

$$V_B = V_G - V_X - V_{FB} - \sqrt{2q\varepsilon_s N_{AA} V_B}/C_o - Q_N/C_o \tag{682.6}$$

$$\approx V_G - V_X - V_{FB} - \sqrt{2q\varepsilon_s N_{AA} V_B}/C_o \tag{682.6A}$$

$$\equiv V_G - V_X - V_{FB} - V_{GTX} \tag{682.6B}$$

Equation (682.6A) can be solved for V_B to give

$$V_{GTX} \equiv \sqrt{2q\varepsilon_s N_{AA} V_B}/C_o \equiv \sqrt{2V_{AA}V_B} \tag{682.7}$$

$$= V_{AA}\{\sqrt{1+[2(V_G-V_X-V_{FB})/V_{AA}]} - 1\} \tag{682.8}$$

$$\approx \sqrt{2V_{AA}(V_G-V_X-V_{FB})} \tag{682.8A}$$

where

$$V_{AA} \equiv q\varepsilon_s N_{AA}/C_o{}^2 = 13.92(N_{AA}/10^{16})(x_o/100A)^2 \text{ mV}. \tag{682.8B}$$

The approximation in (682.8A) is valid since V_{AA} is only about 14 mV as indicated in (682.8B) while V_G-V_X-V_{FB} can be 5 to 20V.

The explicit solution of the energy band bending at the source, $V_B(0)$, as a function of V_G given above can now be used in the electron charge equation to give the drain current. To write Q_N in terms of $V_B(0)$, the Q_N integral given by (682.3) must be evaluated analytically. Let us use the Boltzmann approximation

$$N(x,y) = n_i \exp\{(q/kT)[V_I(x,y)-V_N(x,y)]\} \tag{682.9}$$

and

$$N(x,0) = n_i \exp\{(q/kT)[V_I(x,0)-V_N(x,0)]\} \tag{682.9A}$$

$$= n_i \exp\{(q/kT)[V_I(x,0)-V_X\} \tag{682.9B}$$

then the Q_N integral can be approximated by

$$Q_N(y=0) = q\int_0^{\infty}[N(x,y) - N_B]dx \tag{682.10}$$

$$\propto q\int_0^{1}[N(X,y) - N(1,y)]dX\cdot x_c \tag{682.10A}$$

$$= qN_B\cdot x_c\cdot\exp\{q[V_B + V_X - V_S]/kT\} \tag{682.10B}$$

$$= Q_{N0}\cdot\exp\{q[V_B + V_X - V_S]/kT\} \tag{682.10C}$$

where x_c is an **effective subthreshold electron surface channel thickness**. It can be derived using the depletion assumption $V_B(x) = V_B[1-(x/x_d)]^2$ in (682.10).

$$x_c = x_d\{1 - \exp(-qV_B/kT)[(qV_B/kT) - 1]\}/\{2qV_B/kT\} \tag{682.11}$$

$$\propto x_d/4 \qquad \text{(in error by 4/3)} \qquad V_B \ll kT/q \tag{682.11A}$$

$$\propto L_D\sqrt{kT/2qV_B} \qquad V_B > 4kT/q. \tag{682.11B}$$

L_D is the extrinsic Debye length defined by

$$L_D = \sqrt{\varepsilon_s kT/q^2N_{AA}} = 408.8(T/300)^{½}(10^{16}/N_{AA})^{½}\cdot 10^{-8}\text{cm}. \tag{682.11C}$$

Q_{N0} is the electron charge density at flat band, $V_{I0}=0$, which is given by

$$Q_{N0} = qN_Bx_c \propto q(n_i{}^2/N_{AA})\sqrt{kT/2qV_B}\cdot L_D \qquad V_B>2kT/q \tag{682.12A}$$

$$= 6.573\text{x}10^{-3}\ (10^{16}/N_{AA})^{3/2}\sqrt{1/V_B}(T/300)(n_i/10^{10})^2\ (q/\text{cm}^2) \tag{682.12B}$$

$$= 1.053\text{x}10^{-21}(10^{16}/N_{AA})^{3/2}\sqrt{1/V_B}(T/300)(n_i/10^{10})^2\ (\text{C/cm}^2). \tag{682.12C}$$

At the strong inversion threshold, $V_{I0}=2V_F$ or $V_B(0)=V_{SX}+2V_F$, we have

$$Q_N(V_{I0}=2V_F) = 6.573\text{x}10^{9}(T/300)^{½}(N_{AA}/10^{16})^{½}\ (q/\text{cm}^2). \tag{682.13}$$

The numerical values given above use $\varepsilon_s=11.7\varepsilon_v$ for Si and $\varepsilon_o=3.9\varepsilon_v$ for SiO_2 where $\varepsilon_v=8.854\text{x}10^{-14}$F/cm. Substituting the $Q_N \ll C_oV_G$ approximation of V_B given by (682.6A) into (682.10C), then

$$Q_N(y=0) \propto Q_{N0}(0)\cdot\exp\{q[V_G-V_S-V_{FB}(0)-V_{GTX}(0)]/kT\}. \tag{682.14A}$$

A similar derivation can be carried out to give

$$Q_N(y=L) \propto Q_{N0}(L)\cdot\exp\{q[V_G-V_D-V_{FB}(L)-V_{GTX}(L)]/kT\}. \tag{682.14B}$$

Subthreshold Drain Current Equation

If the MOST is uniform or symmetrical, then $V_{FB}(y=L)=V_{FB}(y=0)$ and $Q_{N0}(y=L)=Q_{N0}(y=0)$. The subthreshold drain current from (682.3) is

$$I_D = (Z/L)D_n[Q_N(y=0) - Q_N(y=L)] \qquad (682.3)$$

$$\simeq I_{D0}\cdot\{\exp[q(V_{GS}-V_{FB}-V_{GTX})/kT]\}\cdot\{1-\exp[-q(V_{DS})/kT]\}. \qquad (682.15)$$

The pre-exponential current coefficient in the above solution is given by

$$I_{D0} \equiv (Z/L)(1/V_B)^{1/2}D_n\cdot Q_{N0} \qquad (682.16)$$

$$= 1.053\text{x}10^{-20}V_B{}^{-1/2}I_{DN1}\ \text{A} = 1.089\text{x}10^{-20}V_B{}^{-1/2}I_{DN2}\ \text{A}. \qquad (682.16B)$$

where $V_B{}^{1/2}\equiv 1$ as a first approximation. The normalization factors are defined by

$$I_{DN1} = (Z/L)(D_n/10)(10^{16}/N_{AA})^{3/2}(T/300)^{1/2}(n_i/10^{10})^2 \qquad (682.16C)$$

and

$$I_{DN2} = (Z/L)(\mu_n/400)(10^{16}/N_{AA})^{3/2}(T/300)^{3/2}(n_i/10^{10})^2. \qquad (682.16D)$$

At the strong-inversion threshold, $V_{I0}=2V_F$, the subthreshold I_D from (682.15) is

$$I_D(V_{I0}=2V_F) = 10.53I_{DN1}\ \text{nA} = 10.88I_{DN2}\ \text{nA} \qquad (682.17)$$

since the electron density is increased by $\exp(2qV_F/kT)=(N_{AA}/n_i)^2$.

Dependence of Subthreshold Current on Drain Voltage

The dependence of the subthreshold drain current on the drain voltage V_D comes from $Q_N(V_D)$. When V_D-$V_S>(2kT/q)+(V_{GTS}$-$V_{GTD})$, $Q_N(V_D)$ can be dropped. Then, the channel current is independent of V_D and completely controlled by the source, gate and substrate voltages. However, it does not saturate at a finite drain voltage like the super-threshold range. Instead, it saturates asymptotically like a reverse-biased p/n junction. Thus, the subthreshold drain 'saturation' current from (682.15) for $V_{DS}\equiv V_D$-$V_S>4kT/q\simeq 100$mV is given by

$$I_D(\text{sat}) \simeq I_{D0}\cdot\exp[q(V_{GS}-V_{FB}-V_{GTX})/kT]. \qquad (682.18)$$

Dependence of Subthreshold Current on Gate Voltage

Equation (682.18) shows that the subthreshold drain current varies with V_{GS} sub-exponentially because V_{GTX} increases with increasing gate voltage as indicated by (682.8) or (682.7) which, when using (682.6B) for V_B, can be written as

$$V_{GTX} = C_o{}^{-1}\sqrt{2q\varepsilon_s N_{AA}[V_{GS}+V_{SX}-V_{FB}-V_{GTX}]}. \qquad (682.19)$$

This shows that there are two cases where the $\sqrt{V_{GS}}$ dependence of V_{GTX} drops out so that I_D is nearly proportional to $\exp(qV_G/kT)$: (1) very thin oxide or large $C_o=\varepsilon_o/x_o$ so that V_{GTX} from (682.19) is much smaller than V_{GS}-V_{FB} in (682.18); and (2) large bulk charge from applying a large reverse substrate/source junction bias, $V_{SX} >> V_{GS}-V_{FB}-V_{GTX}$ in (682.19), so that $V_{GTX} \propto \surd[V_{SX}-V_{FB}-V_{GT}] \neq f(V_{GS})$. These are shown in Fig.682.1(b) by the $V_S-V_X=0V$ and 10V curves on the left for $x_o=100A$, and the $V_S-V_X=5V$ curve on the right for $x_o=1000A$.

For these two cases (thin oxide and large reverse V_{SX}), the gate voltage swing needed to decrease the substrate current by one decade is given by $S \equiv [(d/dV_G)\log_{10}I_D]^{-1} \propto [(d/dV_G)\log_{10}\exp(qV_G/kT)]^{-1} = (\log_e 10)(kT/q) = 2.303(kT/q) = 59.5mV/decade \propto 60mV/decade$ where two additional assumptions are made: the V_G dependence of D_n and x_c (682.11) can be dropped compared with that of the minority carrier (electron) concentration at the SiO_2/Si interface, $N_S(source) \equiv N(x=0,y=0) = (n_i^2/N_{AA})\exp[q(V_{GS} - \ldots)/kT]$. Thus, it takes a gate voltage change of 60mV to decrease the subthreshold drain current by one decade. This is a very important circuit design criterion for cutting off the MOST sufficiently in order to reduce the signal and d.c. leakage and the standby power dissipation.

The gate voltage range of this nearly true exponential dependence is $2V_F = 2(kT/q)\log_e(N_{AA}/n_i) \propto 0.7V$ and it spans $(N_{AA}/n_i)^2 = 10^{12}$ or 12 decades of current. Figure 682.1(b) shows that between the subthreshold (flat band) and intrinsic surface ($V_S=V_F$) points, the subthreshold current is so small that it is masked by leakage and noise currents. However, between intrinsic (V_F) and strong inversion ($2V_F$), there is still an exponential dependence over a 350mV gate voltage range and $\log_{10}(N_{AA}/n_i) \propto 6$ decades of current which is ample for measurements.

The smaller slope or larger voltage swing due to gate voltage drop through the thicker oxide and low or zero substrate bias can be illustrated by

$$S = [(d/dV_G)(\log_{10}I_D)]^{-1} \equiv (dV_{IO}/dV_G)^{-1}\cdot[(d/dV_{IO})(\log_{10}I_D)]^{-1}$$
$$\propto [1 + (C_{it}+C_s)/C_o]\cdot 2.303(kT/q) \qquad (682.20)$$

where the exact differential relationship for dV_G/dV_{IO} is derived from Fig.410.1(d) or (410.7) where $dV_S \equiv dV_{IO}$. It is given by: $dV_G/dV_{IO} = (dV_G/dQ_G)/(dV_{IO}/dQ_G) = [1/C_g]/[dV_{IO}/(-dQ_{IT}-dQ_S)] = [C_o^{-1}+(C_{it}+C_s)^{-1}]+[(C_{it}+C_s)^{-1}] = 1+[(C_{it}+C_s)/C_o]$. Equation (682.20) shows that the gate-voltage swing, S, is greater than $\log_{10}(kT/q) = 2.303(kT/q) \propto 60mV/decade$ for the two reasons already given: small C_o or thick oxide and large C_s or thin Si surface space-charge layer (small or no substrate reverse bias), but also a **third cause**: high C_{it} from large interface trap density which can be used to measure the interface trap density, as shown in section 683.

Temperature Dependence of the Subthreshold Current

The temperature dependence of the subthreshold drain current arises from the same Boltzmann factor that gives the voltage dependences. Thus, the thermal activation energy of the subthreshold current, ΔE, is very large. Including E_G from n_i^2, then $I_D = K\cdot\exp\{(q/kT)[(V_{GS}-V_{FB}-V_{GTX}-(E_G/q))]\} \equiv K\exp(-\Delta E/kT)$ so

$$\Delta E = (E_G/q) - (V_{GS} - V_{FB} - V_{GTX}). \qquad (682.21)$$

Since the subthreshold current spans a range of $2V_F=0.714$ eV for Si-nMOST with $N_{AA}=10^{16}\text{cm}^{-3}$ and $E_G=1.206\text{eV}$, ΔE can vary from 1.206eV at subthreshold voltage to 0.849eV at intrinsic surface, and to 0.492eV at the threshold voltage. Thus, the subthreshold current is highly temperature dependent unlike the superthreshold current. The origin of the large temperature dependence is identical to that of the forward current of a p/n junction diode or collector current of a bipolar junction transistor: minority carrier injection by a forward biased n+/p junction. In bipolar junction transistor (BJT) applications, the large temperature dependence is circumvented by operating the BJT as a current amplifier. However, FET's are voltage controlled resistors or transconductance amplifiers. Thus, the large temperature dependence of the subthreshold current in MOST is a serious hindrance for its application in the subthreshold range as a low power device.

683 Effects of Oxide and Interface Traps

Oxide and interface traps are dangling oxygen and silicon bonds. The MOST's current-voltage characteristics are seriously affected by the presence of charged oxide and interface traps. The traps are undesirable because they cause transistors to age or transistor characteristics to shift, distort and drift, and the integrated circuits to fail. But their presence in MOSTs is inevitable. They exist in trace amounts ($<10^{-4}$ atomic monolayer or $<10^{10}\text{cm}^{-2}$) as residues from manufacturing processes, and they are generated after prolonged operation of MOST at high voltages (~5V) which give high electric fields (>5MV/cm). Areally uniform and charged oxide traps will only displace the I_D-V_G in parallel along the V_G axis, however, interface traps will also distort the I_D-V_G curves. This different gate-voltage dependence allows experimental separation of the interface traps from the charged oxide traps. Variation of the oxide trap density from the source to the drain will also distort the I_D-V_G curves. Then, frequency dependence of trapping at the interface traps must be used to separate the interface and oxide traps. The shifts and distortions of the I_D-V_G curves are similar to those of the C_g-V_G curves described in Figs.403.1(c) and 403.2(c). Surface mobility variation with V_G will further distort the I_D-V_G curves. The mathematical analysis of the effects of oxide and interface traps on the (I_D-V_G) characteristics are described in the following subsections. Experimental data can be compared with the ideal trapless theory to help detect the oxide and interface traps and characterize their fundamental properties (density, location, capture and emission rates, and energy levels).

Oxide Traps

If the charged oxide trap density is areally constant in the gate oxide over the channel, the oxide traps will only displace the drain current along the gate voltage axis without distortion. However, if the oxide charge density varies over the gate and channel area, the I_D-V_G curves will also be distorted. Even if the oxide trap density is initially constant or zero, areal inhomogeneous oxide charge can be built up near the drain junction during the operation of MOSTs by channel hot electron

(CHE) injection from Si into the oxide and the capture of the injected electrons by the oxide traps. This electron injection occurs because electrons in the channel are accelerated by the high voltage drop near the reverse biased n+/p junction to kinetic energies greater than the 3.1eV Si/SiO$_2$ potential barrier. Areal inhomogeneous oxide charge can also be built up near the source end of the channel via Fowler-Nordheim tunnel electron injection (FNTEI) from the Si surface into the oxide because of the high oxide field (~5MV/cm) from the full V_G-V_S (~5V) across the very thin gate oxide (100A). Section 682 describes these high voltage and high field effects.

The quantitative effect of charged oxide trap can be analyzed using the definitions of the threshold and flat band voltages, (681.4A) $V_{GST} = V_{GT} - V_S = 2V_F + V_{GTX} + V_{FB}$, and (681.5) $V_{FB} = \phi_{MS} - (Q_{OT} + Q_{IT})/C_o$. The charged oxide trap contribution is given by $-Q_{OT}/C_o$. Q_{OT} is the effective areal density of charged oxide trap density referred to the SiO$_2$/Si interface. Figure 683.1 shows a sketch of a volume density distribution of charged oxide traps $\rho_{OT}(x,y,E,t)$.

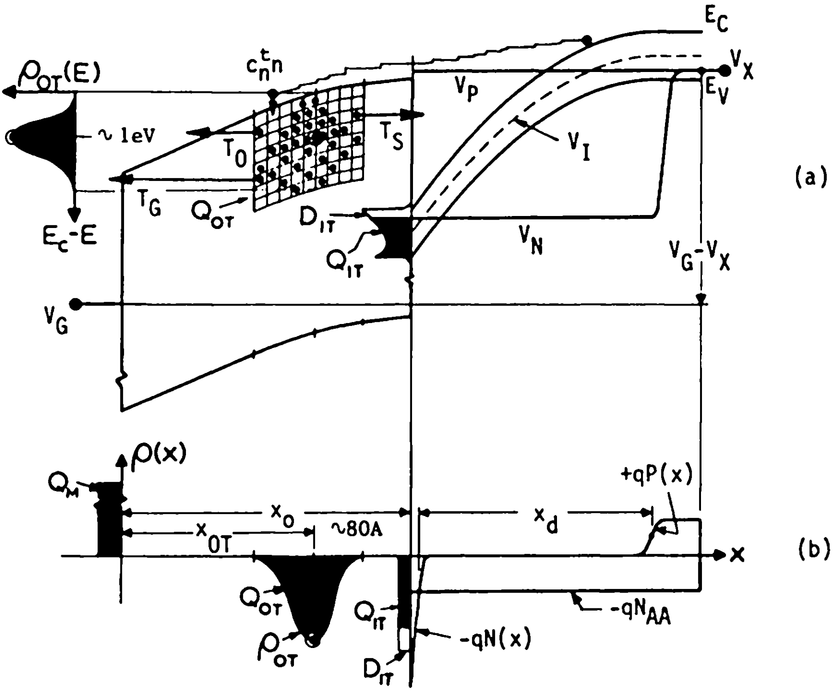


Fig.683.1 A p-Si nMOST showing the effects of charged oxide traps and interface traps. (a) The E-x energy band and density of oxide traps in a state-of-the-art SiO$_2$ discovered by Thompson and Nishida. [Applied Physics Letters 58(12),1262, 25 March 1991.] (b) The $\rho(x)$ charge density with estimated x-distribution of charged oxide traps. E_C-E_T(peak)≈1eV. x_o-x_{OT}≈80A.

An analytical expression can be readily obtained for the oxide charge distribution shown in Fig.683.1 using the parallel-plate capacitance formula of two capacitances in series, one with thickness x and the other (x_o-x). This is given by

$$q_{OT}(y,t) = \int_0^{x_o} (x/x_o)\cdot\rho_{OT}(x,y,t)dx \qquad (683.1)$$

$$= qn_{OT}(y,t)\cdot[x_{OT}(y,t)/x_o]. \qquad (683.2)$$

$n_{OT}(y,t)$ and x_{OT} are the two basic parameters that characterize the charged oxide trap distribution. n_{OT} is the total charged oxide trap density (#/cm^3) defined by

$$n_{OT}(y,t) = q^{-1}\int_0^{x_o} \rho_{OT}(x,y,t)dx. \qquad (683.3)$$

It can be used to measure the charging, discharging and annealing rates of existing oxide traps and the generation rate of new oxide traps. x_{OT} is the centroid of the charged oxide traps measured from the gate/oxide interface and defined by

$$x_{OT}(y,t) = \frac{\int_0^{x_o} x\cdot\rho_{OT}(x,y,t)dx}{\int_0^{x_o} \rho_{OT}(x,y,t)dx}. \qquad (683.4)$$

It can be used to measure the movement and change of shape of the charged oxide trap distribution due to (1) diffusion, drift, generation and annealing of the oxide traps and (2) the space-time dependences of the rate of charging and discharging the existing oxide traps by electrons and holes via tunneling and thermal-optical-Auger-impact transitions (see Table 360.1).

Unfortunately, gate threshold shift measurements from the C_g-V_G and I_D-V_G curves of MOST's and MOSC's can only give the product of these two basic parameters or the effective density of the charged oxide trap, $Q_{OT}=q_{OT}(y,t\to\infty)$ defined by (683.2). This limitation is illustrated by the following ΔV_G formulae.

$$\Delta V_G = V_{GS}(Q_{OT}) - V_{GS}(Q_{OT}=0) \qquad (683.5)$$

$$= - Q_{OT}/C_o \qquad (683.5A)$$

$$= - \int_0^{x_o} x\cdot\rho_{OT}(x,y)dx/\varepsilon_o \qquad (683.5B)$$

$$= - qn_{OT}\cdot x_{OT}/\varepsilon_o = - (qn_{OT}/C_o)\cdot(x_{OT}/x_o). \qquad (683.5C)$$

In order to separate n_{OT} and x_{OT}, a second set of experiments must be performed on the same MOSC or MOST sample in which the measured parameters are dependent on the location or distribution of the charged oxide traps, such as the Fowler-Nordheim tunneling current which depends strongly on the oxide charge variation with x (assuming no variation with y and z). An alternative and easy method is to measure oxides of different thicknesses but one must assume that $\rho_{OT}(x,y)$ is identical at all thicknesses used. Other methods have also been used, such as step oxide etch in which the C-V curve is taken after each oxide thinning step. The derivative dQ_{OT}/dx_o would give volume density of charged oxide traps. The step-etch method is not only destructive but also inaccurate because the oxide etching rate is difficult to control. Scattered particle and sputtering spectroscopic methods have also been used to determine $\rho_{OT}(x,y,t)$ versus depth, x, without much success due to the very low oxide trap density at parts per billion to trillion levels.

Interface Traps

Interface traps will distort the I_D-V_G and C_g-V_G curves because the density of the charge trapped by the interface traps will vary with the electron and hole densities at the SiO_2/Si interface (or with the surface potential and gate voltage as commonly explained which is another defective consequence-consequence rather than cause-consequence explanation). For a cause-consequence example, more electrons will be trapped by the interface traps making Q_{IT} more negative when the electron concentration at the interface is higher.

The general analytical expression of the interface charge density, Q_{IT}, can be derived when the device is not at equilibrium, for example, the MOSC and MOST are pulsed, or there is a large d.c. voltage difference between the drain and the source, resulting in a significant d.c. channel or drain current. The interface traps are assumed to be in equilibrium with the Si at the SiO_2/Si interface, that is, the number of electrons or holes trapped at the interface is determined by the concentration of electrons and holes at the SiO_2/Si interface. To illustrate the analysis technique, we shall assume zero drain-source current, zero steady-state terminal voltages, and an acceptor-like interface trap (neutral when empty and negatively charged when occupied by an electron). Let the traps be distributed in the entire Si energy gap with a density of state $D_{IT}(E_T)$ at each trap energy level E_T as illustrated in the E-x energy band diagram given in Fig.683.1(a). Then the charge density of all the interface traps in the Si energy gap is given by

$$Q_{IT}(P_S,N_S) = -q\int_{E_V}^{E_C} D_{IT}(E_T)f_T(E_T-E_F)dE_T \qquad (683.6)$$

$$= -q\int_{E_V}^{E_C} \frac{D_{IT}(E_T)dE_T}{1+\exp[(E_T-E_F)/kT]} \cdot \qquad (683.6A)$$

D_{IT} is in the unit of ($cm^{-2}eV^{-1}$). $f_T(E_T-E_F)$ is the fraction of the interface traps occupied by electrons at a trap energy E_T, and it is different from the Fermi-Dirac distribution function of band electrons as explained in chapter 3. The positive shift of gate voltage due to negatively charged interface traps is then

$$\Delta V_{GT} = V_{GT}(Q_{IT}) - V_{GT}(Q_{IT}=0) = -Q_{IT}/C_o \tag{683.7}$$

$$= +(q/C_o)\int_{E_V}^{E_C} \frac{D_{IT}(E_T)dE_T}{1+\exp[(E_T-E_F)/kT]} . \tag{683.7A}$$

This result can be used to determine the density of interface traps from experimental data by taking the derivative of the threshold voltage shift with respect to the surface potential or trap energy level E_T, $\partial E_T=-q\partial V_S$. Thus,

$$q(\partial \Delta V_{GT}/\partial E_T) = -\partial \Delta V_G/\partial V_S \tag{683.8}$$

$$= +(q/kTC_o)\int_{E_V}^{E_C} - D_{IT}(E_T)f(E_T-E_F)[1-f(E_T-E_F)]dE \tag{683.8A}$$

$$\simeq -(q/kTC_o)D_{IT}(E_T=E_F)kT. \tag{683.8B}$$

To obtain (683.8A), $\partial f_T(E_T-E_F)/\partial E_T=-f_T(1-f_T)$ is used. To obtain (683.8B), we note that $f_T(E_T-E_F)[1-f_T(E_T-E_F)]$ is a bell-shaped function with a height of 1/4 at $E_T=E_F$ and a width of 2kT. Thus, it can be approximated by a delta function of strength kT with $D_{IT}(E_T)$ evaluated at $E_T=E_F$ with $V_S=2V_F$ (threshold) or any V_S to span the Si energy gap. Equation (683.8B) can compute the density of interface traps from distortions in experimental C_g-V_G or I_D-V_G curves, which is

$$D_{IT}(E_T=E_F) = (C_o/q)[(\partial \Delta V_G/\partial V_S)] \quad (\#\text{state/eV-cm}^2). \tag{683.9}$$

The derivative is taken at constant C_g or I_D from distorted C_g-V_G or I_D-V_G curves.

The density of interface state, D_{IT}, can also be obtained from the subthreshold slope of the I_D-V_G or the subthreshold voltage swing, S, derived in (682.20) which shows that the change of S is proportional to the capacitance of interface trap, C_{it}. From $C_{it}=-\partial Q_{IT}/\partial V_S=-C_o\partial(\Delta V_G)/\partial V_S=qD_{IT}$ using (683.9), then

$$\Delta S = S(q_{IT}) - S(q_{IT}=0)$$

$$= 2.303(kT/q)(C_{it}/C_o) = 2.303(kT/q)qD_{IT}/C_o \tag{683.10}$$

or

$$D_{IT} = (C_o/q)(q/2.303kT)\Delta S \quad (\#/\text{eV-cm}^2). \tag{683.11}$$

This is a particularly powerful and sensitive method since it measures a d.c. current which is much easier to measure than the high-frequency capacitance in the Terman method. However, the change of the subthreshold slope due to mobility reduction from additional scattering by the newly generated interface and oxide traps must be assumed negligible.

684 High Electric Field and High Voltage Effects

High electric field and high voltage effects are known as the **hot electron** and **hot hole effects** or **hot carrier effects**. They arise because the transport parameters (rates of drift-diffusion, generation-recombination, emission-capture, and tunneling) depend on either the electric field or the electric field and applied voltage (or carrier kinetic energy). The fundamental mechanisms underlying these dependencies are described in detail in chapter 3. The drift mobility and diffusivity decrease with increasing electric fields because the rate of carrier energy loss increases due to more frequent collisions with the vibrating Si atoms that create optical phonons. As the electric field and applied voltage increase, the higher carrier kinetic energy causes the carrier generation and emission (detrapping) rates to increase and the carrier recombination and capture rates to decrease. The tunneling rates (interband, band-trap and trap-trap) increase with increasing electric field because of thinner potential barrier.

Thus, high electric field and high voltage can seriously reduce the performance and operating life of semiconductor devices. The degradation of the device characteristics is the principal cause of failure of integrated circuits.

The degradation rate of MOST is more sensitive to high electric fields and high voltages than bulk devices (p/n Junction-Gate FET and BJT) because MOST's geometry and material imperfections. Its geometry has the semiconducting channel or the electrically active layer located right next to a highly heterogeneous interface (the insulator/semiconductor or SiO_2/Si interface). The two material imperfections are: (i) the extremely heterogeneous interface: amorphous insulator on crystalline semiconductor (a-SiO_2/c-Si), and (ii) the amorphous insulator (a-SiO_2) itself. Both of these material imperfections are characterized by disordered and weak (strained or stretched) intrinsic bonds (Si:Si and Si:O) and weak impurity bonds (Si:H and SiO:H). The hydrogen bond energy is roughly 3eV.

Disordered Si intrinsic bonds will give additional random scattering of the electrons and holes, causing mobility reduction. Weak intrinsic and hydrogen bonds can be broken to give Si dangling bonds when hit by energetic electrons and holes or when capturing a thermal electron or hole that releases a significant capture or recombination energy. Weaker bonds can be broken by less energetic electrons or holes which can be accelerated at lower electric fields and voltage. Disordered and weak bonds, after broken, cannot heal or rebond due to the localized displacement of the atom with the dangling bond and its surrounding host atoms. Dangling bonds are electron and hole traps. They can be charged which shift the threshold voltage. Both charged and neutral dangling bonds will scatter electrons and holes but the charged centers will scatter more because of the longer range of the $1/r$ Coulomb force. In the following sections, we shall discuss these effects on the electrical characteristics of MOSTs and obtain analytical I_D-V_D and I_D-V_G equations.

I-V with Electric Field Dependent Mobility

The characteristics of MOSTs are modified when the mobility and diffusivity are dependent on the electric fields. There are two fundamentally different mobility reduction mechanisms which give different field dependences. (i) The longitudinal electric field increases optical phonon scattering rates. (ii) The transverse electric field squeezes or confines the electron to a thinner sheet closer to the imperfect heterogeneous SiO_2/Si interface resulting in increased scattering by interface and oxide traps but decreased scattering by the bulk dopant acceptor ions in the channel depletion layer. In addition, the 3-d bulk phonon scattering is changed to the more effective 2-d surface phonon scattering. These are discussed by Sah and Nishida. [**Properties of Silicon**, Section 17.20, pp.625-635, INSPEC, Institution of Electrical Engineers, London.] The phenomenological field dependences of the mobility are given by

$$\mu = \frac{\mu_0}{[1 + (|E_y|/E_{CL})^{\gamma}]^{1/\gamma}} \qquad \textbf{Longitudinal Electric Field Effects} \qquad (684.1)$$

and

$$\mu = \frac{\mu_0}{[1 + (|E_x|/E_{CT})]^{\delta}} \qquad \textbf{Transverse Electric Field Effects} \qquad (684.2)$$

where μ_0 is the low field mobility, E_{CL} and E_{CT} are the longitudinal and transverse critical electric fields, and γ and δ are empirical parameters which, in theory, vary with both longitudinal and transverse fields and temperature. These five parameters are determined by fitting the simple theory to experimental data. A review of the literature and a summary of the data in oxidized Si are given by Sah and Nishida in the 1988 reference listed above. For convenience, the components of the mobilities are listed below for electrons in the Si inversion surface channel covered by an SiO_2 film. The total low field surface mobility of electron is obtained by adding the reciprocal of these mobility components, via the Matthiessen rule given by (313.12). Data for holes are not available. A rough guide is to reduce the electron mobility by a factor of 3.

$$\mu_{0nLA} = 7.4\text{x}10^5/T \qquad (\text{cm}^2/\text{V-s}) \qquad (684.3A)$$

$$\mu_{0nLO} = 1.0\text{x}10^8/T^{1.9} \qquad (\text{cm}^2/\text{V-s}) \qquad (684.3B)$$

$$\mu_{0nOX} = 10^3(3\text{x}10^{11}/N_{OX})(T/80) \qquad (\text{cm}^2/\text{V-s}) \qquad (684.3C)$$

$$\mu_{0nDP} = 100(N_{DP}L_{DP}^2)^{-1}(T/100)^{3/2} \qquad (\text{cm}^2/\text{V-s}) \qquad (684.3D)$$

$$\mu_{0nSR} = 1.5\text{x}10^{29}q^2/(|Q_B| + |Q_N|)^2 \qquad (\text{cm}^2/\text{V-s}) \qquad (684.3E)$$

$$\mu_{0nCI} = 8.5\text{x}10^8/(N_{ION})^{1/3} \qquad (\text{cm}^2/\text{V-s}) \qquad (684.3F)$$

The temperature, T, is in Kelvin (K). N_{OX} is the charge density of the oxide and interface traps in cm^{-2}. N_{DP} is the dipole density in cm^{-2}. L_{DP} is the dipole length. Q_B and Q_N are the bulk and inversion charge density in q/cm^2. N_{ION} is the ion concentration (cm^{-3}) in the surface space-charge layer, $x=0$ to x_{sc}, and is approximated by $N_{AA}+N_{DD}$ if a compensating shallow donor is present.

The **critical longitudinal electrical field**, E_{CL}, is given by $E_{CL} = \theta_{sat}/\mu_0 = 10^7/\mu_0$ where μ_0^{-1} is the sum of μ^{-1} given in (684.3A) to (684.3F). For example, if $\mu_0=400cm^2/V\text{-}s$, then $E_{CL}=10^7/400=25kV/cm$. The field-dependent exponent γ is approximately 2.0.

The **critical transverse electric field**, E_{TC}, is approximately 100kV/cm. E_x in (684.2) is the peak transverse electric field at the SiO_2/Si interface and δ varies from 0.5 to 2 as inversion increases and is set at 1.0 for this illustration.

Although the longitudinal and transverse electric field dependences occur simultaneously in a MOST, they are treated separately to isolate their individual effects on the I_D-V_D and I_G-V_D characteristics in the following two subsections.

Longitudinal Field Dependent Mobility Effects

The longitudinal electric field dependent mobility, (684.1) with $\gamma=2$, can be substituted into (641.1C) or (643.1) which is the approximate MOST differential equation assuming that the mobility has a longitudinal-field dependence and no transverse-field dependence. Using the following indefinite integral,

$$\int x\sqrt{(x^2-a^2)}dx = \tfrac{1}{2}\{x\sqrt{(x^2-a^2)} - a^2\log_e[x+\sqrt{(x^2-a^2)}]\}, \qquad (684.4)$$

the analytical solution is given by the following transcendental equation

$$\begin{aligned} 2I_{DN}E_{CL}L = {} & + V_{GS}(V_{GS}^2 - I_{DN}^2)^{1/2} - V_{GD}(V_{GD}^2 - I_{DN}^2)^{1/2} \\ & - I_{DN}^2\cdot\log_e\left[\frac{V_{GS} + (V_{GS}^2 - I_{DN}^2)^{1/2}}{V_{GD} + (V_{GD}^2 - I_{DN}^2)^{1/2}}\right] \end{aligned} \qquad (684.5)$$

where $I_{DN}=I_D/(\mu_0C_oZE_{CL})=I_D/(C_oZ\theta_{sat})$. This reduces to the low field equation, (643.2), if the applied longitudinal electric field is small compared with the critical longitudinal electric field, E_{CL}. Generally, even if the average applied field in the channel is smaller than the critical field, near the drain end of the channel, the electric field will rise rapidly to exceed the critical value. However, since the control of the drain current by the gate voltage occurs near the source end of the channel, the low-field/high-field boundary can be estimated by the average channel or longitudinal electric field. Thus, the low-field condition is $V_D/L << E_{CL}$. At $V_D=5V$ and a low field mobility of 400 $cm^2/V\text{-}s$, the low field approximation is valid when the channel length is longer than $L >> V_D/E_{CL} = 5/25kV/cm = 2\mu m$.

Thus, state-of-the-art MOSTs with one-micron and submicron channels will have a large longitudinal electric field dependence.

Although (684.5) predicts the I_D-V_D characteristics and it can be solved numerically, let us consider the analytical solution at very high electric fields in very short channels. Instead of obtaining the asymptotic solution of (684.5) letting E_{CL} be very small or I very large, the solutions can be readily obtained from the original equation (643.1) using high field limiting mobility, $\mu_n(dV/dy) = \theta_{sat} = \text{constant}$:

$$I_D = C_o Z \theta_{sat}(V_G - V_{GT} - V_S) \tag{684.6A}$$

and

$$g_m = C_o Z \theta_{sat}. \tag{684.6B}$$

The current, (684.6A), may be compared with the low-field, constant-mobility solution in the current saturation range, $I_D = (\mu_0 C_o Z/2L(V_G - V_{GT} - V_S)^2$, to give the general high field criterion: $I_D(\text{high-}E_y) < I_D(\text{low-}E_y)$ or

$$E_{CL} \cdot L < (V_G - V_{GT} - V_S)/2.$$

Thus, for $V_{G-}V_{GT}-V_S = 5V$ or $V_D = 5V$ and $E_{CL} = 25kV/cm$, then $L < 1\mu m$ which is consistent with the low-field constant-mobility criterion which gave $L > 2\mu m$.

The key result of velocity saturation is that the drain current is saturated at a lower value than the drain-charge depletion value without velocity saturation. This lowers the transconductance. And it also makes the transconductance constant independent of the gate voltage, as indicated by (684.6B), instead of parabolic. The MOST then becomes a linear but still amplifying device.

Transverse Field Dependent Mobility Effects

The transverse electric field dependences of the mobility in the channel were empirically determined as a function of the maximum Si transverse electric field at the SiO_2/Si interface. They all originate from the geometrical confinement of the electrons into an increasingly thinner sheet closer to the SiO_2/Si interface as the transverse electric field increases. It is not from an electric-field dependent scattering or energy loss rate because the transverse electric field is orthogonal to the electron drift velocity so that there can be no input power to heat up the drifting electrons from the transverse electric field: $\mathbf{F} \cdot \mathbf{v} = -qE_y \cdot v_x = 0$. This crucial physics and simple mechanics has been missed by some book authors and researchers who have published data on mobility reduction as a function of transverse electric field.

The effect of confinement on the MOST I-V characteristics can be analyzed approximately using the simple empirical formulae (684.2) which lumps all the transverse-field dependences due to confinement or inversion layer thickness into one critical or threshold transverse electric field, E_{CT}, and which assumes that the dependences are scaled by the peak transverse Si electric field at the SiO_2/Si interface. At extremely high transverse fields, interfacial atomic roughness (surface roughness) scattering given by (684.3E) must be taken into account, which has an inverse-square transverse field dependence with $\delta=2$ and is readily observed at low temperatures. A more exact analysis including all scattering mechanisms listed by (684.3A) to (684.3F) would require the evaluation of the drain current integral $q\int\mu_n(x,y)N(x,y)dx$ without making the approximation $\simeq\mu_n(x=0,y)q\int N(x,y)dx =\mu_n Q_N$. Thus, the dependences of $\mu_n(x)$ on x must be known which can be derived from quantum mechanical perturbation theory of electron scattering. These x-dependences arise from the x-dependent volume density of the scatters as well as the 3-d to 2-d transition of the phonon spectra and electron energy band structure.

Because of the simple approximation, $\mu_n(x,y)\simeq\mu_n[E_x(x=0,y)]$, the transverse electric field dependences can be analyzed using E_x-dependent or $C_o[V_G-V_{GT}-V(y)]$-dependent multipliers in μ_0 before the y integration is undertaken. Consider the mobility reduction due to 3-d to 2-d phonon confinement given by (684.3A) and (684.3B), then $\delta=1$ in (684.2). Following the same algebraic procedure which led to (684.5), the I-V equation for 2-d phonon scattering with the assumption of $Q_B\neq f(y)$ is (W=Z=channel width=gate width)

$$I_D = \mu_0(W/L)\int_{V_S}^{V_D} Q_N dV/\{1 + [(Q_N+Q_B)/Q_{CT}]\} \qquad (684.7)$$

The **critical transverse gate charge and voltage**, V_{CT}, are related by

$$Q_{CT} = \varepsilon_s E_{CT} = C_o V_{CT}. \qquad (684.7A)$$

$$I_D = \mu_0 C_o(W/L)(Q_{CT}/C_o^2)\left[C_o V_{DS} - Q_{CT}\log_e\frac{C_o(V_{GS}-V_{GTS})+Q_{CT}}{C_o(V_{GD}-V_{GTD})+Q_{CT}}\right] \qquad (684.8)$$

$$\simeq \mu_0 C_o(W/L)\tfrac{1}{2}[(V_{GS}-V_{GTS})^2-(V_{GD}-V_{GTD})^2] \quad \text{if } Q_{CT}\gg Q_N+Q_B \qquad (684.8A)$$

$$\simeq \mu_0\cdot(W/L)Q_{CT}(V_{DS}) \quad \text{if } Q_{CT}\ll Q_N+Q_B. \qquad (684.8B)$$

At drain current saturation, $Q_N(V_{DS}) = C_o(V_{GS}-V_{GTS}-V_{DS})=0$ which gives $V_{DS} = V_{GS} - V_{GTS}$, then the above result becomes the following if we assume that Q_{GTS} and Q_{CT} are constant along the channel length.

$$I_D = \mu_0 C_o(W/L)(Q_{CT}/C_o^2)\left[C_o(V_{GS}-V_{GTS}) - Q_{CT}\log_e\frac{C_o(V_{GS}-V_{GTS})+Q_{CT}}{Q_{CT}}\right] \qquad (684.9)$$

$$\simeq \mu_0 C_o(W/L)\tfrac{1}{2}(V_{GS}-V_{GTS})^2 \quad \text{if } Q_{CT}\gg Q_N+Q_B \qquad (684.9A)$$

$$\simeq \mu_0\cdot(W/L)Q_{CT}(V_{GS}-V_{GTS}) \quad \text{if } Q_{CT}\ll Q_N+Q_B. \qquad (684.9B)$$

Using $E_{CT}=100\text{kV/cm}$ for 2-d phonon scattering and $x_o=10\text{nm}$, then

$$Q_{CT} = \varepsilon_s E_{CT} = 6.466\text{x}10^{11}\ \text{q/cm}^2 = 0.1036\ \mu\text{C/cm}^2 \tag{684.10A}$$

$$E_{OX} = (\varepsilon_s/\varepsilon_o)E_{CT} = 300\ \text{kV/cm} \tag{684.10B}$$

$$V_{CT} = Q_{CT}/C_o = 0.300(x_o/10)\ \text{V}. \tag{684.10C}$$

When the applied gate voltage is above V_{CT}, i.e. $V_{GS}-V_{GTS} > V_{CT}$, mobility reduction due to transverse electric field or confinement becomes important. The above numerical values give a critical gate voltage of 0.300V. Such a small value shows that mobility reduction due to increasing confinement at higher transverse electric fields is important for just about all practical operating conditions. Note that the drain current at saturation again changes from the parabolic dependence on the gate voltage, (684.9A), at low transverse electric fields to the linear dependence, (684.9B), at high transverse electric fields.

Consider a second channel confinement effect on mobility, the increased scattering by atomic roughness of the SiO_2/Si interface. This dominates at very strong inversion or high transverse electric field and at low temperatures when the low-field mobility is very high. Using the Matthiessen rule, $1/\mu_n = 1/\mu_0 + 1/\mu_{SR}$, the surface roughness mobility formulae given by (684.3E) in the drain current equation (643.1), and the indefinite integral

$$\int (x-a)dx/(1+x^2) = (1/2)\log_e(1+x^2) - \text{atan}^{-1}x, \tag{684.11}$$

then, the analytical solution of the drain current assuming $Q_B \neq f(V)$ is

$$I_D = (\mu_0/c_o)(W/L)\int_{V_D}^{V_S} Q_N dQ_N/\{1 + [(Q_N+|Q_B|)/Q_{SR}]^2\} \tag{684.12}$$

$$= \mu_0 C_o(W/L)(Q_{SR}/C_o)^2\cdot$$

$$\left[\frac{1}{2}\log_e\frac{Q_N(V_S)+|Q_B|+Q_{SR}}{Q_N(V_D)+|Q_B|+Q_{SR}} + \frac{Q_B}{Q_{SR}}\tan^{-1}\frac{[Q_N(V_S)-Q_N(V_D)]/Q_{SR}}{1+Q_N(V_S)Q_N(V_D)Q_{SR}^{-2}}\right] \tag{684.13}$$

$$\approx \mu_0 C_o(W/L)\cdot\tfrac{1}{2}[Q_N^2(V_S)-Q_N^2(V_D)]/C_o^2 \qquad \text{if } Q_N+Q_B \ll Q_{SR} \tag{684.13A}$$

$$\approx \mu_0 C_o(W/L)(Q_{SR}/C_o)^2\cdot\log_e\frac{Q_N(V_S)+|Q_B|+Q_{SR}}{Q_N(V_D)+|Q_B|+Q_{SR}} \qquad \text{if } QN+Q_B \gg Q_{SR}. \tag{684.13B}$$

At drain current saturation, $Q_N(V_{DS}) = C_o(V_{GS}\text{-}V_{GTS}\text{-}V_{DS}) = 0$ which gives $V_{DS} = V_{GS} - V_{GTS}$, the preceding results then become

$$I_D = \mu_0 C_o (W/L)(Q_{SR}/C_o)^2$$

$$\cdot\left[\frac{1}{2}\log_e\frac{Q_N(V_S)+|Q_B|+Q_{SR}}{0\quad+|Q_B|+Q_{SR}}+\frac{Q_B}{Q_{SR}}\tan^{-1}\frac{Q_N(V_S)+0}{Q_{SR}\quad+0}\right] \qquad (684.14)$$

$$\simeq \mu_0 C_o (W/L)\cdot \tfrac{1}{2}[Q_N(V_S)/C_o]^2 \qquad \text{if } Q_N+Q_B \ll Q_{SR} \qquad (684.14A)$$

$$\simeq \mu_0 C_o (W/L)(Q_{SR}/C_o)^2\cdot\log_e\frac{Q_N(V_S)+|Q_B|+Q_{SR}}{0\quad+|Q_B|+Q_{SR}} \qquad \text{if } QN+Q_B \gg Q_{SR}. \qquad (684.14B)$$

The critical gate charge, critical oxide electric field at $\mu_0 = 1000\text{cm}^2/\text{V-s}$, and critical gate voltage at $x_o = 10\text{nm}$ for surface roughness scattering to become important are given by

$$Q_{SR} = q\sqrt{1.5\times10^{29}/\mu_0} = 1.22\sqrt{\mu_0/1000} \qquad (10^{13}q/\text{cm}^2) \qquad (684.15)$$

$$Q_{SR} = 1.95\sqrt{\mu_0/1000} \qquad (\mu\text{C/cm}^2) \qquad (684.15A)$$

$$E_o = Q_{SR}/\varepsilon_o = 5.66\sqrt{\mu_0/1000} \qquad (\text{MV/cm}) \qquad (684.15B)$$

$$V_{SR} = Q_{SR}/C_o = E_o x_o = 5.66\sqrt{\mu_0/1000}(x_o/10) \qquad (\text{Volt}). \qquad (684.15C)$$

Four conclusions can be made. (i) The electric field at which surface roughness scattering becomes important (5MV/cm) is much larger than that of 2-d phonon scattering (100kV/cm). (ii) Surface roughness scattering is important in the operation of submicron MOST at the high end of the gate bias voltage ($V_{GS}\text{-}V_{GTS} \rightarrow 5\text{V}$) and oxide electric field (5 MV/cm). (iii) Surface roughness scattering is increasingly important as the mobility is higher, thus, it is important at low temperatures which is the condition under which most of the surface roughness data were reported. (iv) Q_{SR} at $\mu_0 = 1\text{cm}^2/\text{V-s}$ is $3.9\times10^{14}(q/\text{cm}^2)$ which is nearly equal to the areal density of the Si-O bond at the SiO_2/Si interface. This is not an accidental coincidence: it is consistent with the randomness of the Si-O bond length and angle because SiO_2 is noncrystalline or amorphous. The very large critical electric field or very tight confinement (5A or bond length) indicates a very-short-range scattering potential (about a bond length) which can only be detected when the electrons are squeezed into the interfacial layer to within a bond distance from the random bonds. This is again consistent with the short range of the random Si-O bonds in the SiO_2/Si interfacial layer.

Electric Field and Voltage Dependent Generation-Recombination-Trapping

Current-voltage characteristics of a MOST can change during operation at high electric fields and high voltages due to charging, discharging and generation; and also annihilation via thermal annealing and hydrogenation (including chlorination, fluoridation and other bond formations) of the oxide and interface traps and shallow energy-level dopant impurities in the Si surface space-charge layer. The density change of the oxide traps and the dopant impurities will shift the threshold and subthreshold voltages through the two terms $(Q_{OX}+Q_B)/C_o$ or V_{FB} and V_{GTX}. The density change of the interface traps will distort the I_D-V_G transfer characteristics through the term Q_{IT}/C_o.

The fundamental energy exchange mechanisms involving electrons and holes were already discussed in 360 and listed in Table 360.1. The additional mechanisms involve hydrogenation and de-hydrogenation of the defect and impurity traps because proton (1 fermi or 10^{-13}cm radius) and hydrogen atom (0.53A radius) are very mobile in SiO_2 and Si even at room temperatures due to their small size, and because they are very reactive or have high rates of association and dissociation with these **protonic** traps due to their low binding energy. The electric field dependences arise from injecting the electrons or holes into the oxide by tunneling through a thin SiO_2/Si potential barrier whose thickness and hence the tunneling rate depends on the electric field, or by avalanche injection over the potential barrier whose rate depends on the density of the energetic electrons from acceleration by the electric field in the Si surface space-charge layer. The voltage dependences arise from the minimum kinetic energy of impacting electron or hole required to break the weak intrinsic Si-Si or Si-O bonds and the hydrogen bonds. The hydrogen bond energy (about 3.5eV) is the lowest among all atomic bonds. Review of the mechanisms [684.1] and data [684.2] were given recently.

[684.1] C.T.Sah, "Models and Experiments on Degradation of Oxidized Silicon," Solid-State Electronics,13(2), pp.147-167, Feb.1990.

[684.2] C.T.Sah, et.al., 20 sections in Chapter 17, "Insulating Layers on Silicon Substrate," in **Properties of Silicon**, INSPEC, IEE-London, 1988.

Other small but less mobile metallic impurities (Li, Na and K) act like hydrogen and proton in SiO_2 but are generally not present in today's ultra clean gate oxide and Si. However, it is not possible to avoid hydrogen since it is present in most chemicals and gases used in the fabrication of Si transistors and integrated circuits. Chlorine and fluorine (also bromine and iodine) will also bond with the defect and impurity traps to deactivate them electronically just like hydrogen. Cl and F bonds are more stable than the H bond because they are stronger and because the larger and heavier Cl and F atoms are not as mobile as H. Recent (1990) experiments seem to suggest lower oxide and interface trap generation rate at high electric fields in Cl-doped and F-doped SiO_2.

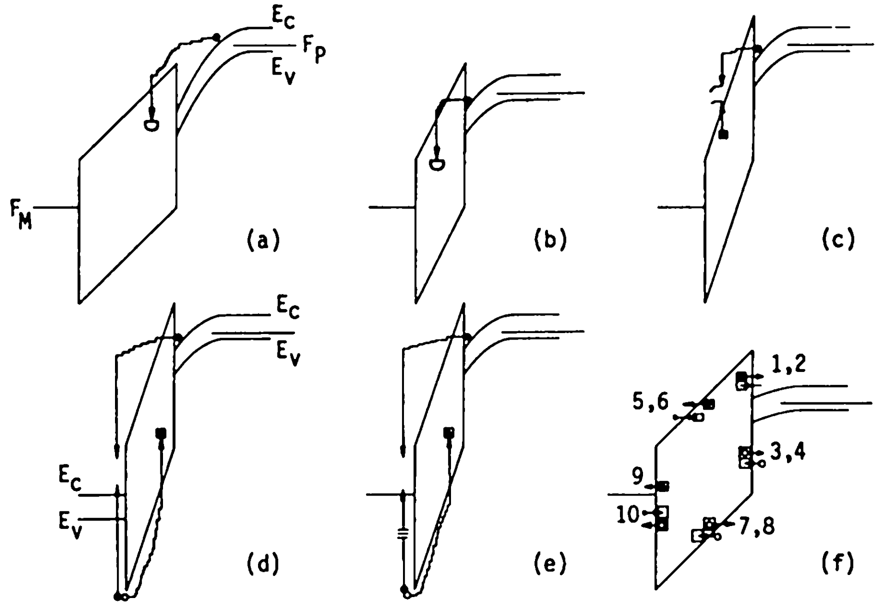


Fig.684.1 Charging and discharging of oxide traps by electron and hole capture and emission transitions. (a) Capture of AEIed electron. (b) Capture of FN-TEIed electron. (c) Impact emission of trapped electron by AEIed or FN-TEIed and oxide-field accelerated electron. (d) Capture of electron impact generated gate-conductor hole. (e) Capture of plasmon-decay generated gate-conductor hole. (f) Ten elastic electron tunneling transitions into or out of an oxide trap.

Five Trap Charging-Discharging Processes via Capture and Emission

Five multi-step charging-discharging processes of oxide traps are shown in Figs.684.1(a) to (e). (a) is avalanche electron injection (AEI) followed by electron capture by an oxide trap. (b) is Fowler-Nordheim tunneling electron injection (FN-TEI) followed by electron capture by an oxide trap. (c) is acceleration of AEIed or TEIed electron by the oxide electric field to impact threshold followed by electron impact emission (EIE). (d) is acceleration of AEIed or TEIed electron by a very high oxide voltage drop to kinetic energies of at least 2.4 ± 0.6eV for Al or p-Si gate or 1.2 ± 0.6eV for p-Si gate to overcome the 4.3 ± 0.6eV gate-conductor/SiO_2 hole barrier (assuming $E_G(Si)=1.2$eV and $E_G(SiO_2)=8.6\pm0.6$eV); followed by: electron impact generation (EIG) of an thermal-electron/energetic-hole pair in the Si- or metal-gate, injection of the hole into the oxide, transport of the hole in the high oxide electric field, and hole capture either by an oxide hole trap or a weak intrinsic or hydrogen interface bond, the latter resulting in the generation of an interface trap. (e) is similar to (d) except that EIG is replaced by: surface plasmon generation (SPG) by a more energetic electron (8.55eV for Al/SiO_2 and 9.45eV for Si/SiO_2), and decay of the surface plasmon into electron/energetic-hole pair.

Ten Trap Charging-Discharging Processes via Elastic Tunneling

Ten single-step direct or elastic tunneling processes are shown in Fig.684.1(f). These are electron or hole tunneling transitions to or from the oxide trap into the Si and SiO_2 conduction or valence band and the gate metal. For Si-gate, transitions 9 and 10 are replaced by 1 to 4.

The existence and dominance of most of these are experimentally demonstrated. The first two, AEI in (a) and FN-TEI in (b) have been measured by many researchers and engineers from different laboratories. EIE in figure (c) and tunnel electron emission (TEE) of transition (5) in figure (f) have been measured in detail by Thompson and Nishida recently (1990). SPG of figure (e) has been demonstrated using metal gate by Fischetti while EIG has been suggested by Sah, Sun and Tzou (1985) using Si gate.

Eight Trap Generation-Annihilation Processes via Hydrogenation

Two generation and deactivation processes of oxide and interface traps via hydrogenation-dehydrogenation processes are shown in four figures, Figs.684.2(a)-(d). The four generation processes can also occur at weak intrinsic Si-Si or Si-O bonds. The first two, (a) and (b), are oxide trap generation and interface annihilation (also possible generation). The second two, (c) and (d), are interface trap generation and oxide trap annihilation (also possible generation). The steps of each of the four processes are described in the following paragraphs.

(a) Oxide trap generation via: FN-TEI or AEI, electron acceleration by a high oxide electric field and high oxide voltage to a high kinetic energy greater than the hydrogen bond energy, breaking a hydrogen bond in oxide, released hydrogen migrating away towards the SiO_2/Si interface; and interface trap hydrogenation by the released and arrived hydrogen.

(b) Oxide trap generation via: FN-TEI or AEI, electron acceleration by a high oxide electric field and voltage to a high kinetic energy, electron/energetic-hole generation in the gate conductor via EIG or SPG-EHG, hole injection into oxide, hole capture by hydrogen bond in oxide, released hydrogen migrating away towards the SiO_2/Si interface; and interface trap hydrogenation by the arrived hydrogen.

(c) Interface trap generation by: the energetic AEIed electron impact via breaking a hydrogen bond at SiO_2/Si interface; released hydrogen migrating into the oxide from the interface, and oxide trap hydrogenation by the released interfacial hydrogen.

(d) Interface trap generation via: FN-TEI or AEI, electron acceleration in high oxide electric field and voltage, electron/energetic-hole generation in gate conductor via EIG or SPG-EHG, hole injection into oxide, hole capture by hydrogenated interface trap, released hydrogen migration away from SiO_2/Si interface; and oxide trap annihilation by the released interfacial hydrogen.

In the last step of each of the four processes, instead of hydrogenating or annihilating a trap, the released hydrogen could also generate a trap by breaking a strained intrinsic Si:Si or Si:O bond to give a hydrogenated and electronic inactive bond and a dangling electronic active bond. Three possibilities are:

$$\mathrm{H\cdot + Si{:}Si(strained) \longrightarrow Si{:}H + Si\cdot} \qquad (684.16A)$$

$$\mathrm{H\cdot + Si{:}O(strained) \longrightarrow SiO{:}H + Si\cdot} \qquad (684.16B)$$

and

$$\mathrm{H\cdot + Si{:}O(strained) \longrightarrow Si{:}H + SiO\cdot} \ . \qquad (684.16C)$$

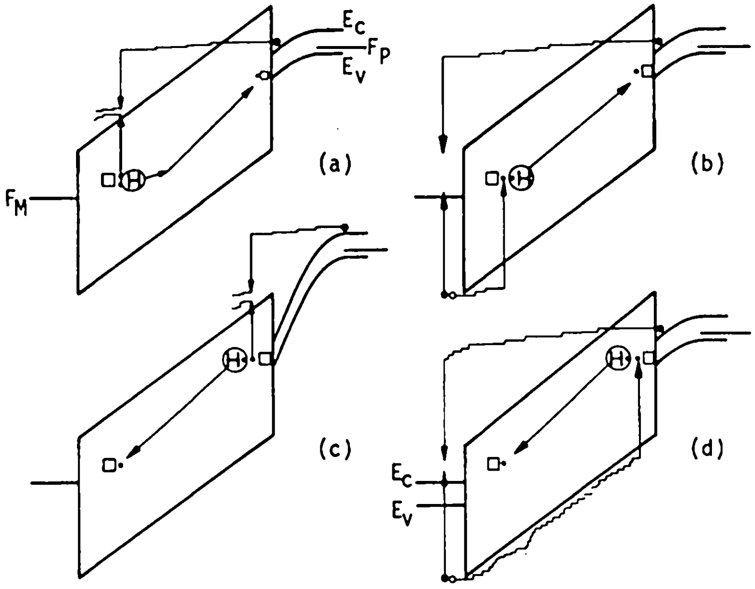


Fig.684.2 Generation and annihilation of oxide and interface traps by hydrogen reactions. For explanation of each step, see paragraphs in text labeled (a) to (d).

Hydrogenation of Boron and Group-III Acceptors in Si

This effect was discovered in 1972 by Sah and Fu in p-Si MOSC when they observed a dramatic decrease of the strong-inversion high-frequency capacitance after exposure to 1 Coulomb/cm^2 of 10keV electrons shown in Fig.403.2(d). The experiment was repeated in 1982 by Sah, Sun and Tzou using AEI and FN-TEI in p-Si MOSC who then gave the correct atomic model: boron acceptor hydrogenation. More than one hundred follow-up experiments were published by American, European and Asian researchers subsequent to the Sah-Sun-Tzou initial announcement. These have verified the correctness of the atomic model given by

Sah-Sun-Tzou's original article which included the detailed position of the hydrogen in the hydrogenated boron acceptor center and the chemical formulae

$$\mathrm{H\cdot} + \mathrm{Si_3{:}{:}{:}B{:}Si{:}{:}{:}Si_3^-} + h^+ \longrightarrow \mathrm{Si_3{:}{:}{:}B\cdot H\cdot Si{:}{:}{:}Si_3}. \tag{684.17}$$

This hydrogenation reaction occurs in the Si surface space-charge layer of the SiO_2/Si interface. Hydrogenation of the other group-III acceptors (Al, Ga and In) was also first reported by Sah, Sun and Tzou in 1983. The hydrogen source can be: the released hydrogen by AEI and FN-TEI like Figs.684.2(a) and (b); by heating a aluminum gate which contains much H and OH; and by exposure of the p-Si surface to an atomic hydrogen source such as a hydrogen plasma.

This is obviously an important threshold voltage instability mechanism since hydrogen is always present in VLSI manufacturing and the reduction of active boron acceptor concentration by hydrogen will shift the threshold voltage negatively.

The boron acceptor hydrogenation reaction has been modeled by Sah (see section 17.17 of [684.2]) with the following three equations whose rate coefficients are given in Table 684.1 using the g-r-t-t notation of chapter 3.

$$e^* + \mathrm{Si_3{:}{:}{:}B{:}Si^-} + \mathrm{H\cdot} \longleftrightarrow \mathrm{Si_3{:}{:}{:}B\cdot H\cdot Si} + e^* + e^- \tag{684.18A}$$

or

$$h^+ + e^* + \mathrm{Si_3{:}{:}{:}B{:}Si^-} + \mathrm{H\cdot} \longleftrightarrow \mathrm{Si_3{:}{:}{:}B\cdot H\cdot Si} + e^* \tag{684.18B}$$

$$e^* + \mathrm{H\cdot} + \mathrm{H\cdot} \longleftrightarrow \mathrm{H{:}H} + e^* \tag{684.19}$$

$$e^- + h^+ \longleftrightarrow 0 \tag{684.20}$$

Table 684.1
Rate Constants of Acceptor Hydrogenation in Si at 300K

PARAMETER	Unit	B	Al	Ga	In
AEI Voltage	V	60	65	65	65
AEI Frequency	kHz	50	300	300	300
Oxide Thickness	A	1000	1000	900	800
N_{AA}	$10^{16}cm^{-3}$	4.08	0.633	0.652	0.799
A(0)	$10^{16}cm^{-3}$	4.07	0.605	0.644	0.799
$A(\infty)$	$10^{16}cm^{-3}$	0.25	0.069	0.165	0.465
Initial Jump	$1-[A(0)/N_{AA}]$	0.005	0.044	0.012	0.005
Initial Delay	t_d s	310	460	1000	2500
Initial Delay	F_d $10^{17}e/cm^2$	0.32	1.5	25	-
Cross Section	σ_H $10^{-20}cm^2$	6.72	4.47	1.23	3.80
Rate Constant	k_H $10^{-6}s^{-1}$	8.9	8.7	5.3	2.4
H emission rate	e_H $10^{-6}s^{-1}$	0.54	0.95	1.34	1.4
H capture rate	$c_H H_G$ $10^{-6}s^{-1}$	8.3	7.8	4.0	1.0
H release rate	e_X $10^{-6}s^{-1}$	3200	2200	1000±500	400±100

685 Short Channel and Narrow Gate Effects

The gate width, Z, in the MOST differential equation (641.1), (641.1C) and (643.1) and the channel length, L, during the integration of this equation to give (643.2), are assumed to be geometrical constants determined by the lithographic mask and fabrication processing. In general, Z and L of the surface inversion layer that provides the output (drain) current are not constants but are dependent on the drain and gate voltages. The voltage dependences become increasingly important when the geometrical width and length of the MOST decrease towards 0.1 micron or 100nm. They are also important in designing Si MOSTs for high-voltage MOS integrated circuits. The two geometrical factors each influence three electrical characteristics of the MOST. **Short channel** (i) increases the drain conductance by preventing the drain current from saturating, (ii) reduces the threshold voltage, and (iii) increases the transconductance. **Narrow gate** (i) decreases the drain current, (ii) increases the threshold voltage, and (iii) increases the drain current at high gate voltage. The physical origins are illustrated by figures and described in the following paragraphs. The physical models and figures are then used to derive a simple analytical solution for each effect.

Three Short Channel Effects and Lowly Doped Drain

The three short channel effects are: (i) finite (nonzero) saturation drain conductance and nonsaturating drain current, (ii) reduction of the gate threshold voltage, and (iii) higher transconductance. (i) is due to the encroachment of the drain-junction space-charge layer into the channel at high applied drain voltages which decreases the **electrical channel length**. (ii) is due to the encroachment of the source-junction space-charge layer into the channel which reduces the substrate or body charge density controlled by the gate voltage. (iii) is due to insufficient phonon scattering of electrons during transit through extremely short channels, enabling the electrons to accelerate to velocities higher than the steady-state saturation value, Θ_{sat}, known as **velocity overshoot**.

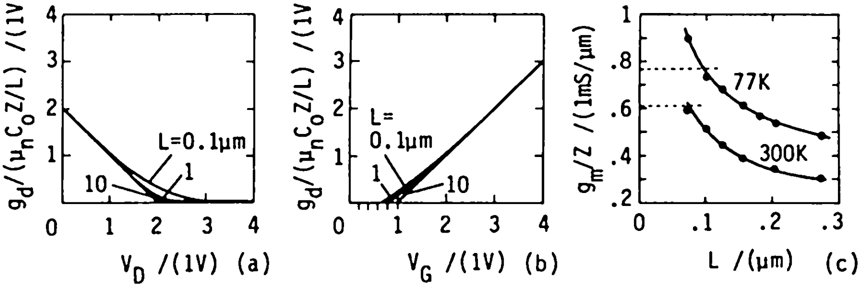


Fig.685.1 The three short channel effects on electrical characteristics. (a) g_d-V_D showing $g_{dsat}>0$. (b) g_d-V_G showing smaller V_{GT}. (c) g_m-L showing $g_m(\text{data})>C_o\Theta_{sat}Z=(----)$ [685.1].

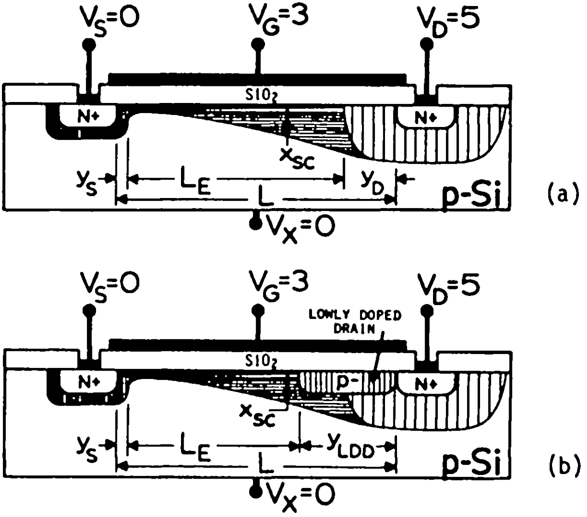


Fig.685.2 The first short channel effect: the nonsaturation drain conductance. **(a)** The normal MOST. **(b)** The MOST with lowly doped drain (LDD) to reduce nonsaturation.

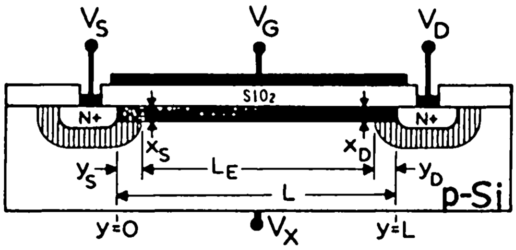


Fig.685.3 The second short channel effect (**Yau effect**): the threshold voltage reduction.

The three short channel effects on the electrical characteristics are illustrated in Figs.685.1(a), (b) and (c). Figure (a) shows that the drain conductance does not drop to zero in the drain current saturation range as predicted by the simple theory shown in Fig.653.3. Figure (b) shows that the threshold voltage is lowered. Figure (c) shows that the transconductance rises above the expected maximum from velocity saturation predicted by (684.6B).

Physical shortening of the channel is shown in Fig.685.2(a) to help illustrate the origin and derive a simple mathematical model. The dark stripe at the SiO_2/Si interface is the inversion layer. The three shaded regions are the depleted space-charge regions. The middle space-charge region shaded by the horizontal lines is mainly controlled by the gate voltage. The left and right space-charge regions shaded by the vertical lines are mainly controlled by the source and drain voltages respectively. The channel length in the MOST equation (to be denoted by L_E while reserving L for the geometrical length) is the distance between the edge of the source ($y=y_{SOURCE}\equiv y_S$) and drain ($y=L-y_{DRAIN}\equiv y_D$) space-charge layers at the interface $x=0$. At the drain end, the inversion electron density reaches zero, $Q_N=0$, and the depleted drain-junction space-charge region begins. L_E is known as the **electrical** or **effective channel length.** As indicated in figure (a), it is less than the geometrical channel length or the distance between the source and drain junctions at the SiO_2/Si interface, $L_E = L-y_D-y_S < L$.

In the simple 1-d MOST theory, the drain depletion point ($y=y_S+L_E$) is determined by $Q_N(y=L_E)=0$ or $V_{DS} = V_{GS}-V_{GTS} \equiv V_{DSS}$. From Fig.685.2(a), it is obvious that the physical origin of channel shortening is the thickening of the depleted drain-junction space-charge layer with increasing applied drain voltage. This is identical to the mechanism that causes base thickness modulation or thinning in bipolar junction transistor discovered by Early in 1952, known as the **Early Effect** (section 738). Figure 685.2(a) at once gives a simple solution

$$L_E = L - y_D - y_S \tag{685.1}$$

$$= L - \sqrt{2\varepsilon_s[V_{bid}+V_{DS}-(V_{GS}-V_{GTS})]/qN_M} - \sqrt{2\varepsilon_s(V_{bis}+V_{SX})/qN_{MS}} \tag{685.1A}$$

$$\simeq L - \sqrt{2\varepsilon_s[V_{bid}+V_{DS}-(V_{GS}-V_{GTS})]/qN_M} \tag{685.1B}$$

where $y_D\equiv y_{DRAIN}$ and y_S are the voltage dependent space-charge-layer thickness of the n+/p drain and source junctions respectively at the SiO_2/Si interface. Equations (685.1) and (685.1A) apply when $V_{DS} > V_{GS}-V_{GTS}-V_{bid}$ in the drain-current saturation range. The analytical solution, (685.1A), is obtained using (531.1) of the space-charge-layer thickness of an abrupt n+/p junction with constant N_{AA} and N_{DD} where the mean concentration is $N_M = N_{AA}N_{DD}/(N_{AA}+N_{DD})$. This simple one-dimensional analytical solution was first obtained by Reddi and Sah at Fairchild Semiconductor in 1961. If a reverse d.c. voltage is also applied between the source/substrate n+/p junction, then L is further reduced by $y_S\equiv y_{SOURCE}$ as indicated by (685.1A). Since $I_D \propto 1/L_E$, (685.1A) indicates that the drain current does not saturate and the drain conduction is not zero when the drain-source voltage becomes greater than $V_{GS}-V_{GTS}-V_{bid}$. The saturation drain conductance is easily obtained as follows using $I_D \propto L_E^{-1}$.

$$g_d = \partial I_D/\partial V_{DS} = I_D\cdot\partial\log_e I_D/\partial V_{DS} = -\ I_D\cdot\partial\log_e L_E/\partial V_{DS} \tag{685.2}$$

$$= (I_D/V_{DS})[y_D/(L-y_D)] \qquad (\text{for } V_{DS}>V_{DSS}\equiv V_{GS}-V_{GTS}) \tag{685.2A}$$

where

$$y_D = \sqrt{2\varepsilon_s[V_{bid}+V_{DS}-(V_{GS}-V_{GTS})]/qN_M} \tag{685.2B}$$

Equation (685.2A) shows that the saturation drain conductance can be as high as 1/2 of the average drain conductance, I_D/V_{DS}, if the electrical channel is shortened to 1/2 of the geometrical channel, $L_E=L-x_D=L/2$, or half of the channel is encroached by the drain junction space-charge layer, $x_D=L/2$. As V_{DS} increases beyond $V_{GS}-V_{GTS}$, the drain conductance will decrease at a rate slower than $V_{DS}^{-1/2}$.

To reduce or eliminate the channel shortening effect, a lowly doped (p^-) channel region next to the drain/channel junction is incorporated as indicated in Fig.685.2(b). The length y_{LDD}, thickness x_{LDD}, and acceptor concentration $N_{AA\text{-}LDD}$ of this lowly doped region are selected such that the excess drain voltage, $V_{DS}\text{-}(V_{GS}\text{-}V_{GTS})$, is confined to drop through its length by making the drain-junction space-charge-layer punch-through or exceed the length of this layer. This is known as the **Lowly Doped Drain** (LDD). It was first introduced (~1980) into the design of very short channel Si MOSTs by Japanese engineers. However, its principle was invented by Early in 1954 for PNIP and NPIN bipolar junction transistors to withstand high collector-base voltages (see section 738) and has been well known to bipolar junction transistor engineers for more than 25 years.

Threshold voltage reduction, (ii), in short channel MOST is due to the encroachment of the source-junction space-charge layer into the channel so that a part of the bulk charge is controlled by the source-junction field and not by the gate voltage. This source-junction-controlled bulk charge region is the area shaded by the square mesh next to the source n+/p junction shown in Fig.685.3. This effect was first analyzed in 1972 by Leo D. Yau of Intel using the charge-control method when he was at the Bell Telephone Laboratories. This **Yau effect** is increasingly important for very short channels because it reduces the gate threshold voltage by a very significant amount. In Yau's analysis, a geometrical division, based on electrostatics or anticipated solution of Poisson's equation, was made to subtract out the part of the substrate or body charge which is controlled by the source-junction field instead of gate field. This reduces the gate-controlled bulk charge near the source junction, $N_{AA}x_Sy_S$ of (681.2), and hence reduces the threshold voltage due to the two body effects, V_{GSTX}. The geometrical-graphical construction of Fig.685.3 immediately gives the reduction of the threshold voltage due to the smaller bulk charge at the source/channel junction. By inspection, it is

$$\Delta V_{GTS} = -\ qN_{AAS}x_S(y_S/2L)/C_o \tag{685.3}$$

where

$$x_S \propto \sqrt{2\varepsilon_s(2V_{FS}+V_S-V_X)/qN_{AAS}(x=x_S)} \tag{685.3A}$$

and

$$x_{SO} = \sqrt{2\varepsilon_s(2V_{FS}+V_S-V_X)/qN_{AAS}(x=0)}. \tag{685.3B}$$

We let $x_S=y_S$ in this simple analysis using the assumption $N_{AAS}(x=0) \propto N_{AAS}(x=x_{sc}) = N_{AAS}$ = constant at the source junction. In this ideal case of constant N_{AAS}, the threshold voltage reduction (685.3) is simplified to

$$\begin{aligned}\Delta V_{GT} &= -\ \varepsilon_s(2V_F+V_{SX})/(C_oL)\\ &= -\ (\varepsilon_s/\varepsilon_o)(x_o/L)(2V_F+V_{SX}).\end{aligned} \tag{685.4}$$

For a submicron Si MOST example, $\varepsilon_s/\varepsilon_o \propto 11.7/3.9=3$, $x_o=100\text{A}=0.01\mu\text{m}$, $2V_F=700\text{mV}$, and let $V_{SX}=0$, then

$$\Delta V_{GT}\cdot L = -\ (\varepsilon_s/\varepsilon_o)(x_o)(2V_F+0) \propto -\ 3\text{x}0.01\text{x}700 = -\ 21\text{mV}\cdot\mu\text{m}. \tag{685.5}$$

Thus, the threshold voltage decreases inversely with decreasing channel length at a rate of about 20 mV·μm. This is consistent with submicron MOST data of Fig.685.1(c) [685.1].

The third short channel effect, (iii), is the observed higher transconductance than the theoretical maximum predicted by the steady-state velocity saturation value, θ_{sat}, given by (684.6B), $g_m(exp) > g_m(sat) = C_o Z \theta_{sat}$. Its origin is an insufficient number of optical phonon scattering events experienced by the electron during its drift through the very short channel to maintain the thermal equilibrium (Boltzmann) distribution in energy and velocity due to random collisions. Thus, electrons can be accelerated to velocities greater than θ_{sat}, approaching its Newtonian velocity in vacuum (known as **ballistic transport and velocity overshoot**) when the channel is very short. Figure 685.1(c) demonstrates the higher transconductance from velocity overshoot in 0.1μm Si nMOST's measured at 77 and 300K as recently reported by IBM [685.1]. Series resistances, R_s, from the source and drain layers and contacts limit the accuracy of the intrinsic transconductance, g_{m0}, calculated from the measured extrinsic transconductance, g_m, since R_s reduces the measured extrinsic transconductance by $g_m = g_{m0}/(1 + g_{m0}R_s)$.

[685.1] See the excellent overview on short channel effects by G.A. Sai-Halasz, M.R.Wordeman, D.P.Kern, S.A.Rishton, E.Ganin, T.H.P.Chang and R.H.Dennard, "Experimental technology and performance of 0.1-μm-gate-length FETs operated at liquid-nitrogen temperature," IBM J.Research & Development 34(4), pp.452-465, July 1990.

Three Narrow Gate Effects

There are three effects from narrowing the gate width: (i) higher threshold voltage due to fringe field, (ii) lower channel current near the threshold due to a narrower inversion layer than the gate electrode, $W_E < W_G$, and (iii) higher drain current and conductance at large gate voltage due to a wider inversion channel than the geometrical gate, $W_E > W_G$. They are numerically labeled on the g_d-V_G curve in Fig.685.4 and their physical origins illustrated by Figs.685.5(c) and (d).

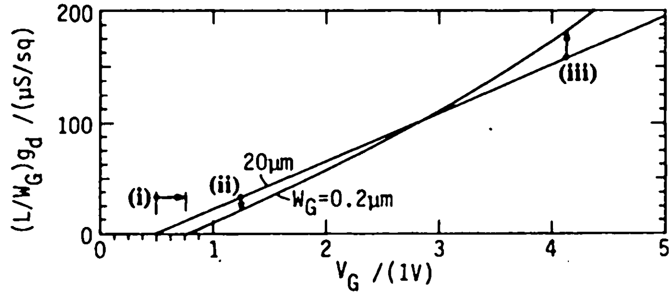


Fig.685.4 The effects of narrow gate on the threshold voltage and drain conductance. The theory was computed numerically by C.R.Ji from solving the 2-d Poisson equation for an Si nMOST with $W_G = 0.2\mu m$, $N_{AA} = 1.7\times10^{16} cm^{-3}$ and $x_o = 400A$. [From IEEE Trans. ED-30, 635(1983).]

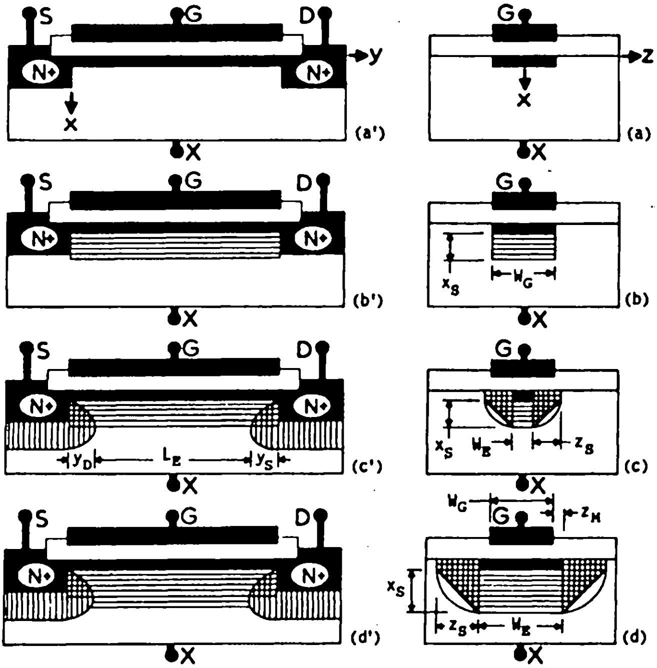


Fig.685.5 Cross-sectional views of a MOSFET in the length direction (primed figures on the left) showing the short channel effects, and in the width direction (unprimed figures on the right) showing the narrow gate effects. Darkened area of gate, drain, source, and inversion channel depicts high electron concentration. Areas shaded by horizontal lines are 1-d space-charge regions. Areas shaded by both horizontal and vertical lines are regions encroached by fringe-fields whose area is controlled by the drain and source voltages but not the gate voltage. (a) Ideal. (b) Real 1-d. (c) Real 2-d at low V_G. (d) Real 2-d at high V_G. [From C.T.Sah, Proc. of International Symposium on VLSI Technology, Systems and Applications - Taipei, 165-169(1983).]

The higher threshold voltage due to narrower gate, denoted by (i) in Fig.685.4, can be readily computed from the increased bulk charge from the two edge areas shaded by square mesh, ΔQ_B, in Fig.685.5(c). This increased bulk charge must be overcome by additional gate charge or gate voltage before an inversion channel can be induced. Thus, the threshold-voltage increase is

$$\begin{aligned}\Delta V_{GT} &= +\,2(|\Delta Q_B|/C_o)(z_S/2W_G) = +\,2(qN_{AAS}x_S/C_o)(z_S/2W_G)\\ &= +\,2(\varepsilon_s/\varepsilon_o)(x_o/W_G)(2V_F+V_{SX}). \qquad (685.6)\end{aligned}$$

where $z_S = x_S = \sqrt{[2\varepsilon_s(2V_F+V_{SX})/qN_{AAS}]}$ is used. Per unit length, this increase of the threshold voltage is twice as large as the decrease due to short channel given by (685.4), because at the source junction there are two sides of the channel in the width direction and only one end in the length direction. If the cross-sectional area of the two edges is approximated by a pi (quarter circle) instead of the triangle used in (685.6), then the edge bulk charge would be $\pi/2$ larger, increasing the threshold voltage given in (685.6) by $\pi/2$. Let $x_o = 100$A, $2V_F = 700$mV, and $V_{SX} = 0$, then

$$\Delta V_{GT}\cdot W_G = 2(\varepsilon_s/\varepsilon_o)(x_o)(2V_F+V_{SX}) \qquad (685.7)$$

$$\approx 2\cdot 3\cdot 0.01\cdot 700 \quad = \ +42\mathrm{mV\cdot\mu m}. \qquad (685.7A)$$

This is a very significant increase even in the first generation VLSI technology when the gate width was 5 micron wide with ΔV_{GT}=(42/5)x(1000A/100A)=80mV for a 1000A oxide. This would be increased by (12.7/0.7) to 1.4V if V_{SX}=12V as used in the first generation 4k DRAMs to give enhancement mode nMOSTs.

The lower drain conductance and the distortion of I_D-V_G curve near the higher threshold voltage in narrow gates, labeled (ii) in Fig.685.4, can also be derived using Fig.685.5(c). It shows that the center line of the channel begins to conduct first and the width of the line or inversion layer (electrical channel), W_E, is zero at the higher threshold voltage. W_E then increases rapidly towards the geometrical width, W_G, as V_G is increased. The missing inversion charge of the narrower channel is taken up by the fringe bulk charge at the two edges, thus,

$$Q_N(W_G-W_E) = 2|Q_B|z_S/2 = qN_{AAS}x_Sz_S \qquad (685.8)$$

$$= qN_{AAS}x_S{}^2 \qquad (685.8A)$$

$$= 2\varepsilon_s[2V_F-V_X+V(y)] \qquad (685.8B)$$

giving

$$Q_NW_E = Q_NW_G - 2\varepsilon_s[2V_F-V_X+V(y)]. \qquad (685.9)$$

The drain current near the threshold region, $I_D = \mu_nQ_NW_E(dV/dy)$ is then

$$\int_0^L I_Ddy = I_DL = \int_{V_S}^{V_D}\mu_nQ_NW_E(dV/dy)dy$$

$$= \int_{V_S}^{V_D}\mu_nW_G\left[Q_N(V) - 2(\varepsilon_s/W_G)[2V_F-V_X+V(y)]\right]dV. \qquad (685.10)$$

The second term of the integrand gives the reduction of the drain current which is

$$\Delta I_D = -\mu_n(\varepsilon_s/L)[(2V_F+V_{DX})^2 - (2V_F+V_{SX})^2]. \qquad (685.11)$$

The conductance reduction, (ii) in Fig.685.4, is then

$$\Delta g_d = \partial\Delta I_D/\partial V_D = -2\mu_n(\varepsilon_s/L)(2V_F+V_{DX}). \qquad (685.12)$$

These analytical solutions have not been reported previously. They show that the reduction of drain current and conductance depends on V_D but not V_G.

At large gate voltages, the channel or drain current is higher than the simple 1-d theory of long channel as depicted by (iii) in Fig.685.4. This is due to a wider channel or inversion layer than the geometrical gate. The wider inversion layer is induced by the fringe electric field extending beyond the two edges of the gate electrode. The additional channel current and conductance can also be derived approximately by a simple geometrical construction as follows. Using Fig.685.5(d), the fringe channel width, $z_M = (W_E - W_G)/2$, is derived from a 1-d estimate of the fringe electric field by noting that $Q_N(z=z_M)=0$. Thus,

$$0 = Q_M \qquad + Q_{SS} + Q_S \tag{685.13}$$

$$= C_{oz}(V_G - V) \quad + Q_{SS} - Q_N + Q_B \tag{685.13A}$$

$$= C_{ozM}(V_G - V_S) + Q_{SS} - 0 \; + Q_{BO} \quad \text{at } y=0 \text{ and } V=V_S \tag{685.13B}$$

where

$$C_{oz} = \varepsilon_o/x_{oz} \equiv \varepsilon_o/\sqrt{x_o^2+z^2} \tag{685.13C}$$

$$C_{ozM} = \varepsilon_o/x_{ozM} \equiv \varepsilon_o/\sqrt{x_o^2+z_M^2} \tag{685.13D}$$

$$x_{ozM} = \varepsilon_o(V_G - V_S)/(-Q_{SS} - Q_B) = x_o(V_{GS}/V_{GT}) \tag{685.13E}$$

$$z = \sqrt{x_{oz}^2 - x_o^2} = x_o\sqrt{[(V_G - V)/V_{GT}]^2 - 1} \tag{685.13F}$$

and

$$z_M = \sqrt{x_{ozM}^2 - x_o^2} = x_o\sqrt{[(V_G - V_S)/V_{GT}]^2 - 1}\ . \tag{685.13G}$$

By setting $Q_N(z=z_M)=0$ in (685.13A), then (685.13B) gives the maximum additional channel width which occurs at the source. It is given by (685.13E) where use is made of $C_oV_{GT} = -(Q_{SS}+Q_B) = -Q_{SS}+|Q_B|$.

The additional drain current and drain conductance due to the fringe electric field can be readily obtained by integration.

$$\Delta I_D = 2\mu_n(dV/dy)\int_0^{z_M} Q_N(z)dz \tag{685.14}$$

$$= 2\mu_n(dV/dy)\int_0^{z_M}\left[\frac{\varepsilon_o(V_G - V)}{\sqrt{(x_o^2+z^2)}} + Q_{SS} + Q_B\right]dz \tag{685.15}$$

$$= 2\mu_n(dV/dy)\left[\varepsilon_o(V_G - V)\log_e(z+\sqrt{x_o^2+z^2})\Big|_0^{z_M} + (Q_{SS}+Q_B)z_M\right] \tag{685.16}$$

where z and z_M are given by (685.13E) and (685.13G). Q_B in the fringe regions is assumed to be independent of z. The additional drain conductance from the two edge regions at and above drain current saturation, $V_{DS}=V_D - V_S=0$, is then

$$\begin{aligned}\Delta g_d = \partial\Delta I_D/\partial V_D &= 2(\mu_n/L)[\varepsilon_oV_{GS}\cdot\log_e(z_{MO}+x_{ozM})/x_o + (Q_{SS}+Q_B)z_{MO}]\\ &= 2(\mu_n\varepsilon_o/L)\Big[V_{GS}\cdot\log_e[\sqrt{(V_{GS}/V_{GTS})^2-1}+(V_{GS}/V_{GTS})]\\ &\qquad + V_{GTS}\sqrt{(V_{GS}/V_{GTS})^2-1}\Big]\end{aligned} \tag{685.17}$$

where $V_{GTS}=-(Q_{SS}+Q_B)/C_o=V_{FB}+V_{GTSX}$ is the threshold voltage including substrate bias. This result is close to the exact curve of $W_G=0.2\mu m$ given in Fig.685.4.

A fourth narrow-gate effect, not listed earlier, is the increase of the intrinsic transit time, $C_{g\text{-}edge}/g_{m0}$, from the edge capacitance of the narrow gate.

690 OTHER FIELD-EFFECT TRANSISTORS - EVOLUTION HISTORY

By definition, all field-effect transistors operate on the principle of conductivity modulation by an applied transverse electric field. They can be grouped according to their gate composition, gate geometry, and channel composition. There are: (1) three gate compositions: the metal/semiconductor (M/S) Schottky barrier gate, the metal/insulator/semiconductor (MOS) gate, and p/n junction gate; (2) two gate geometries: one-sided and two-sided gates; and (3) two channel compositions: the induced and doped channels. Invention of the FET's followed closely this grouping. It is described next, using Figs.690.1(a)-(e).

MESFETs

The first transistor ever invented was actually a metal-gate/p-semiconductor-film/glass-substrate ($Al/Cu_2S/SiO_2$) heterojunction field-effect transistor whose cross-sectional view at three bias voltages are shown in Fig.690.1(a). It was invented by Lilienfeld in 1926 in the first of his three patents filed during 1926-1928 and issued during 1930-1933 [690.1]. It was reinvented in 1939 by Shockley at the Bell Telephone Laboratories [690.2]. Carver Mead built the first successful unit at Caltech in 1966 [690.3]. It has been known as the metal-semiconductor field-effect transistor (MESFET). Taking advantage of higher electron mobility in GaAs ($5000cm^2/V$-s) than Si ($1000cm^2/V$-s), MESFETs of submicron channel length in n-GaAs film on insulating GaAs substrate have been built for 100GHz millimeter-wave amplification in the 1980's. Since 1989-1990, production of picosecond small-scale integrated (SSI) digital circuits (1000 gates) has begun using n-channel depletion and enhancement mode (by lower channel doping) GaAs-MESFET inverter for potential applications in supercomputers, high-speed 3-d graphic display, and navigation-control-surveillance systems. A review was given [690.4] on the two IBM research efforts which reported 0.25-μm channel-length GaAs MESFET that achieved $f_{MAX}=60GHz$ and $t_{delay}=17ps$.

MOSFETs

The first oxide gate FET was also invented by Lilienfeld, in his second transistor patent which was filed in 1930. It was a depletion-mode Al-gate/Al_2O_3/p-Cu_2S/glass thin-film transistor (TFT) shown in Fig.690.1(b). {See Fig.2 in [690.1] for the original structure.} Pearson and Shockley attempted to build a Metal-gate/air-gap/Si-film/glass-slide FET structure in 1948 without success due to high concentration of surface states on the bare Si surface under the metal-gate. The first successful modern MOSFET was the enhancement-mode inversion-channel Si n-MOST reported by Kahng and Atalla of Bell Telephone Laboratories in 1960. Their $Al/SiO_2/Si$ structure used a thermally grown oxide whose cross-sectional view at three biases are shown in Fig.690.1(b'). Recent (1989-1990) research interest has focused on the Silicon on Insulator (SOI) thin-film MOSFET structure shown in Fig.690.1(c) using a buried layer of SiO_2. Technologists have

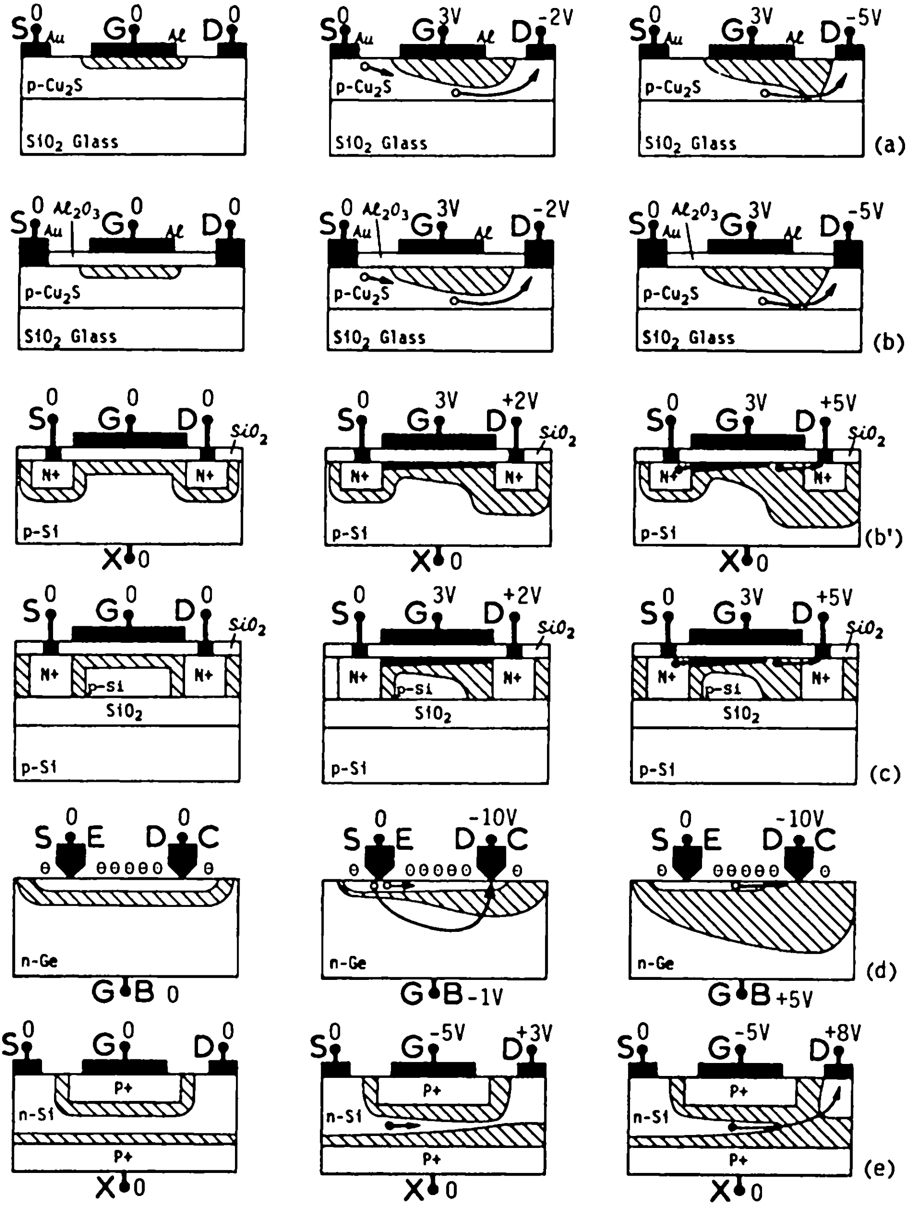

S 0 Au G 0 Al D 0
p-Cu₂S
SiO₂ Glass
S 0 G 3V D -2V
S 0 G 3V D -5V
(a)
Al₂O₃
(b)
N+
SiO₂
p-Si
X 0
D +2V
D +5V
(b')
p-Si
SiO₂
(c)
S E D C
0 -10V
n-Ge
G B 0
G B -1V
G B +5V
(d)
P+
n-Si
G -5V
D +3V
D +8V
(e)

Fig.690.1 Cross-sectional view of five field-effect transistors at three biases. (a) Thin-film MESFET. (b) Doped channel thin-film MOSFET. (b') Modern inversion-channel MOSFET. (c) SOI thin-film MOSFET. (d) The 1948 point-contact surface-ion induced inversion-channel JGFET of Bardeen and Brattain. (e) Bulk JGFET.

been continually challenged to build this structure because it approaches the ideal MOST, having essentially no source and drain junction capacitances and no bipolar coupling via minority carriers between adjacent nMOST and pMOST such as those shown in Fig.672.23(b). Three methods have been used to produce the SOI substrate: (1) evaporate a Si film on an oxidized Si and crystallize the Si film at a high temperature, (2) implant a high concentration of oxygen into the Si surface whose oxygen concentration peaks below the Si surface and form the buried SiO_2 layer by heating at 1000C or higher; and (3) thermally bond the oxide surface of two oxidized Si wafers and chemically etch down one of the Si wafers to a desired thickness (a few micrometers). Method (3) gives the best Si film with essentially no defect but thickness uniformity is difficult to maintain.

The First JGFET (Surface-Ion Induced Inversion-Channel)

For historical interest and possible future application, the surface-ion-induced inversion-channel substrate-junction-gate FET (JGFET) is now described. Its cross-section view at three biases are shown in Fig.690.1(d) for the n-Ge point-contact transistor structure on which Bardeen and Brattain discovered the transistor effect in 1947 [690.5]. The electrode labels are: S for the point-contact source which is also the emitter denoted by E, D for the point-contact drain which is also the collector denoted by C, and G for the substrate gate electrode which is also the base denoted by B. E, C and B are the symbols for the bipolar junction transistor (BJT) described in the next chapter, and they were first used by Bardeen and Brattain to describe the operation of the point contact transistor. The E and C point contacts were made of tungsten whisker. They were thought to be minority carrier (hole) injecting and collecting contacts respectively. The collector point contact attained good rectification and breakdown characteristics after heating by a current pulse to 'form' the contact, probably from diffusing in an acceptor impurity.

There are two current paths for the holes 'injected' from the source contact into the n-Ge when a forward bias ($V_{SG}>0$) is applied to the emitter-base or source-gate junction as depicted by the middle figure of Fig.690.1(d): (i) the long bulk path of minority carriers (holes) through the bulk region of the n-Ge which is the fundamental mechanism of BJT, and (ii) the short surface path through the p-type surface inversion channel between the source and drain contacts. The pre-existing p-type surface inversion channel is induced by the negative ions on the surface and in the interior of the thin germanium oxide film grown from the electrolyte used to treat the Ge surface. The observed I_D-$V_{DS}(V_{GS})$ characteristics, denoted as I_C-$V_{BE}(V_{CE})$ by Bardeen and Brattain, can be accounted for by the channel current rather than the minority-carrier bulk current as follows. Holes are 'injected' by the forward-biased source (emitter) contact, drift through the p-type inversion channel, enter the high-field space-charge layer of the reversely biased collector (drain) junction, and generate many electron-hole pairs by interband impact to give the non-saturating collector (drain) current with increasing collector voltage as observed in the 1948 point-contact transistor characteristics. Thus, the enhancement-mode FET principle can completely account for the point-contact transistor characteristics when the source (emitter) junction is forward biased.

In the depletion mode or when the source-gate (or emitter-base) junction is reverse biased as shown by the right-most figure of Fig.690.1(d), the characteristics would approach that of the conventional FET with drain current saturation. However, if the bulk charge is large and the negative surface ion charge is small, the transistor can be in the cut-off region even at $V_{GS}=0$, i.e. $V_{GTS}<0$. [See Problem 690.3.] In this case, there would be neither channel nor drain (or collector) current at larger reverse bias, such as $V_{GS}=+5$, since V_{GS} must be negative and smaller than V_{GTS} to induce a p-channel. The critical balance between the bulk charge and the surface ion charge necessary to give a built-in inversion channel at zero applied voltages was probably the main cause of irreproducibility and instability (ion drift) in the early point-contact transistors. The surface-ion induced inversion-channel JGFET mode of operation of one of the transistor structures disclosed in the Bardeen-Brattain 1948 transistor patent was not recognized until a recent analysis of the various transistor inventions [690.5]. Thus, strictly, the JGFET was invented by Bardeen and Brattain.

JGFET

The first manufacturable FET was the doped channel p/n junction- gate field-effect transistor (JGFET) shown in Fig.690.1(e). It was invented by Shockley theoretically in 1952 [690.6] to overcome the surface states which caused the failure of his 1948 Metal/air/Si FET experiments with Pearson. A germanium JGFET was built by Decay and Ross in 1953 [690.6] at Bell Labs. The JGFET built on bulk Si (or Ge) does have one advantage over the Si MOSFET. JGFET has smaller $1/f$ and $1/f^2$ low-frequency noise because there is no surface or interface traps to give random fluctuation of trapped charges. But even this advantage is disappearing due to the recent vigorous engineering efforts to lower the interface trap density at the SiO_2/Si interface in order to manufacture low-noise Si MOSFETs and MOS integrated circuit chips for detecting and processing very low intensity infrared and optical signals at low temperatures (77K).

Two Current Saturation Mechanisms (Channel Pinch-off and Carrier Depletion)

The rightmost figures of Fig.690.1(a)-(e) illustrate the two distinct drain current saturation mechanisms often missed by textbook and research authors. In the doped channel FETs, such as the MESFET in (a), MOSTFT in (b), and JGFET in (e), drain or collector current saturation is due to the geometrical **pinch-off** of the thin doped conduction channel to zero thickness near the drain or collector contact. In the induced inversion channel FETs, such as the modern Si MOST in (b'), SOI MOST in (c), and the 1948-Bardeen-Brattain point-contact transistor in (d), drain or collector current saturation is due to **carrier depletion** at the drain or collector junction, not geometrical pinch-off of channel thickness.

High-Mobility Confined-Channel Heterojunction FETs

The intrinsic speed of a field-effect transistor is determined by the mobility and drift velocity of the carrier in the FET channel. Thus, it is desirable to reduce or eliminate the scattering mechanisms in the channel. A high-mobility n-channel Si MOSFET would be obtained if we could remove (i) the acceptor dopant impurity atoms in a thin channel under the oxide and (ii) the interface roughness and oxide charges. In this ideal structure, the electron mobility is completely limited by lattice (acoustic and optical) phonon scattering and may attain a value greater than the empirical surface value computed by (684.3A) and (684.3B) which give 1096 $cm^2/V\text{-}s$ at 300K. It can even be greater than the bulk value 1500 $cm^2/V\text{-}s$. Compared with 300-600 $cm^2/V\text{-}s$ in current state-of-the-art Si nMOSTs, a speed improvement of 2.5X to 5X is attainable. At 77K, the electron mobility in the surface channel would be 9610x26042/(9610+26042) = 7020$cm^2/V\text{-}s$ from (684.3A) and (684.3B), giving a 10x to 20x increase in speed. This was a well-known idea of the 1960's vintage when electron mobility data on oxidized Si surface was obtained. Recently, thin crystalline layers of $Al_xGa_{1-x}As$ with different composition, x, have been grown using the molecular beam epitaxy technique in order to confine the electrons and holes to the pure GaAs layer to achieve extremely high mobilities. Electron mobilities of μ_n=8000(300K) and $2x10^5$(77K) $cm^2/V\text{-}s$, and hole mobilities of μ_p=400(300K) and 6000(77K) $cm^2/V\text{-}s$ have been reported in confined GaAs layers. GaAs FETs with metal or crystalline GaAs gate have been fabricated on these high mobility layers. An excellent overview of the 1990 state-of-the-art of the various high mobility confined-channel heterojunction field-effect transistors (with acronym HFET, HMCCHJ-FET, HMCFET or HCHFET) was given by Kiehl, Solomon and Frank of IBM [690.7]. Figure 690.2 summarizes the cross sectional views and E-x energy band diagrams which illustrate the operation principle. Recent advances have been made by Meyerson at IBM on commensurate growth of Ge_xSi_{1-x} coherently strained film on Si substrate which promise high electron and hole mobilities in confined layers. (See section 77n.) Because of this breakthrough, the decade old IBM research programs on the following GaAs based FET's have been terminated in 1990. Nevertheless, we shall still give a summary because of the many novel device structures and ideas.

Fig.690.2(a) shows the modulation doped structure known as MODFET. The channel is induced by the positively donor impurity in the $Al_xGa_{1-x}As$ 'insulating' film which is analogous to positive charge in the gate oxide of a Si nMOST. The electrons are confined in the 2-d inversion channel on a 0.5μm-thick high mobility p-GaAs film. The GaAs MODFET suffers from severe instability because a deep center (known as the DX center and shown as squares in the E-x diagram) is always present in the $Al_xGa_{1-x}As$ gate-insulator film with a concentration equal to the donor concentration. Threshold voltage instability is induced by electron trapping at this deep DX center.

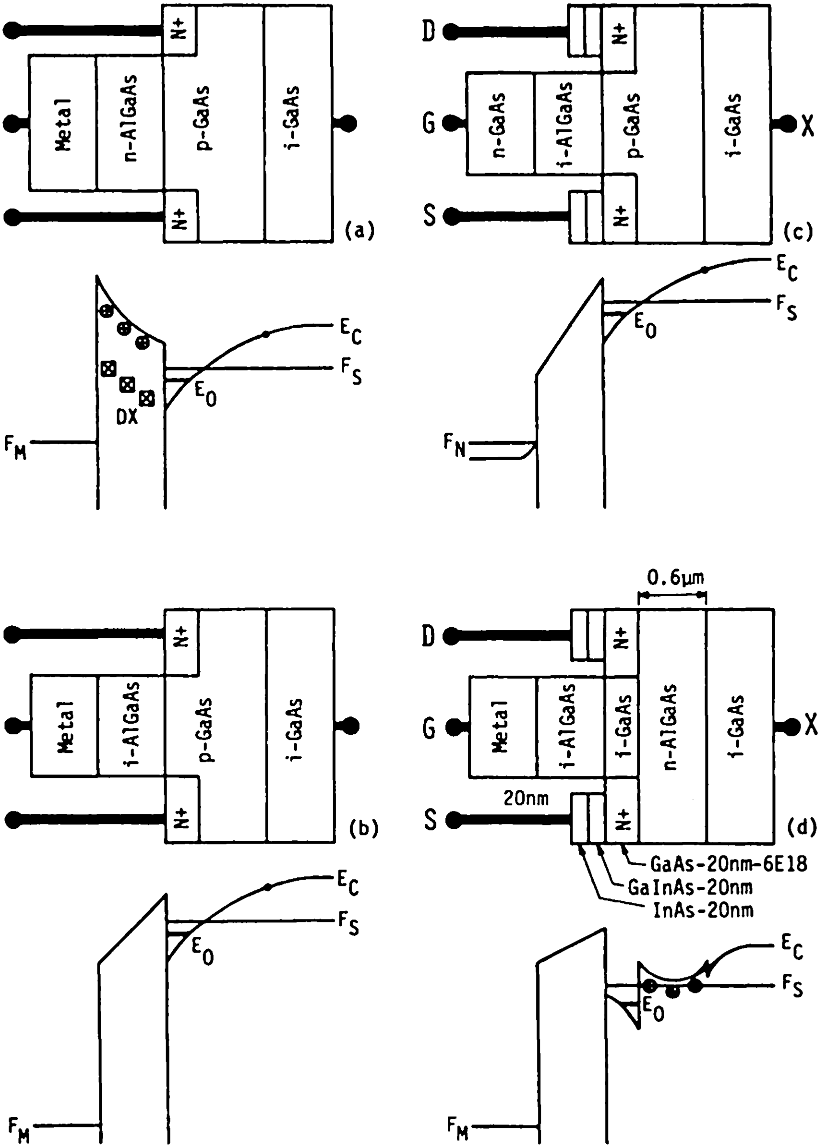

N+
Metal
n-AlGaAs
p-GaAs
i-GaAs
N+
(a)
E_C
F_S
E_0
DX
F_M
D
N+
G
n-GaAs
i-AlGaAs
p-GaAs
i-GaAs
X
S
N+
(c)
E_C
F_S
E_0
F_N
N+
Metal
i-AlGaAs
p-GaAs
i-GaAs
N+
(b)
E_C
F_S
E_0
F_M
0.6μm
D
N+
G
Metal
i-AlGaAs
i-GaAs
n-AlGaAs
i-GaAs
X
20nm
S
N+
(d)
GaAs-20nm-6E18
GaInAs-20nm
InAs-20nm
E_C
F_S
E_0
F_M

Fig.690.2 The cross sectional views and the E-x energy band diagrams of four HMCCHJ-FETs. (a) MODFET. (b) MISFET. (c) SISFET. (d) QW-MISFET.

Fig.690.2(b) shows the gate-voltage induced inversion-channel structure which eliminates the donor in the $Al_xGa_{1-x}As$ gate. This is precisely the same as the inversion-channel Si MOSFET except that the energy gap of $Al_xGa_{1-x}As$ is much smaller than that of SiO_2. The heterojunction FET research community has acronymed this the **MISFET** where **I** is the $Al_xGa_{1-x}As$ gate 'insulator'.

Fig.690.2(c) is the GaAs heterojunction equivalent of the Si MOSFET with a doped poly-Si gate. It was coined **SISFET** by the IBM researchers. This structure was extensively investigated since the threshold voltage can be controlled by doping the GaAs gate. Such a threshold control would facilitate the fabrication of both depletion mode and enhancement mode n-channel and p-channel FETs to build low-power high-speed CMIS (CMOS) circuits.

Fig.690.2(d) shows the next more-complex structure in which the channel is the thin high-mobility intrinsic (pure) GaAs layer between two $Al_xGa_{1-x}As$ layers. The deeper $Al_xGa_{1-x}As$ layer can be doped to control the threshold voltage. This is known as the quantum well metal-'insulator'-semiconductor FET or **QW-MISFET** since the electron channel is confined in a very thin potential well formed by the two $GaAs/Al_xGa_{1-x}As$ heterojunction barriers. The conduction band states in the thin GaAs well is quantized into two-dimensional bands because the x-momentum of the electron (in the thickness direction) is quantized by the thin potential well. The 1-d bound or 2-d ground state is E_0. The other two quantum well structures, **QW-MODFET** and **QW-SISFET**, are obvious.

Speed records measured by ring oscillator (fan-out = 1) as of the third quarter of 1990 were: 5.8ps at 1.8mW in a 0.35μm-gate-length GaAs HMCFET at 77K from ATT; 5.9ps at 30mW in a 0.1μm-gate-length GaAs MESFET at 300K from Japan; and 14ps from a Si nMOST at 77K from IBM [690.7].

Recent breakthrough in commensurate growth of monolayers of Ge_xSi_{1-x} on Si substrate at IBM offers hetero-Ge/Si MOSFET with high electron and hole channel mobilities for CMOS. A possible p-channel structure is poly-Si(gate)/SiO_2/i-Si/i-Ge_xSi_{1-x}(channel)/n-Si(substrate). The holes are kept away from the SiO_2/Si interfacial roughness scattering by the i-Si and confined in the high-mobility pure i-Ge_xSi_{1-x} layer by its valence band discontinuity with the n-Si substrate (see sections 77n).

[690.1] See Fig.1 and description in Chih-Tang Sah, "Evolution of the MOS Transistor - from Conception to VLSI," Proc. IEEE, 76(10), pp.1280-1326, October 1, 1988.

[690.2] See Table 1 and reference 6 of [690.1] on Shockley's MESFET.

[690.3] See Table 2 and reference 98 of [690.1] on Mead's MESFET.

[690.4] T.N.Jackson, et.al., "Submicron-gate-length GaAs MESFETs," IBM J.Res.Develop. 34(4), 495-505, July 1990.

[690.5] See section III.B. on p.1286 of [690.1] on Bardeen-Brattain's JGFET.

[690.6] See Table 1 and references 30 and 34 of [690.1] on JGFET.

[690.7] R.A.Kiehl, P.M.Solomon and D.J.Frank, "Heterojunction FETs in III-V compounds," IBM J.Res.Develop. 34(4), 506-529, July 1990.

699 BIBLIOGRAPHY

There are many MOS circuit books but few MOS device books which give adequate device physics as this chapter. Some circuit books give rather concise and compact analysis of the device characteristics without justifying or even elaborating on the approximations. A selected historical list is given. To supplement and go beyond the mathematical analyses given in this chapter, the reader can follow the sequence with the superscript 'd' for device, 'c' for circuit applications, and 'f' for fabrication. The unscripted references can provide additional information of historical interest.

[699.1]^{d} Robert H. Crawford (Texas Instruments), **MOSFET in Circuit Design**, 146pp, McGraw-Hill Book Co., New York, (1967). This is probably the first complete MOS transistor and circuit textbook. It was written for industrial device and circuit design engineers. In chapters 1-3, it describes all the factors that affect the strong-inversion d.c. characteristics, including substrate doping and surface states, except the subthreshold diffusion current. Switching and small-signal responses of the basic circuit building blocks are analyzed in chapters 4 and 6.

[699.2]^{d} Simon M. Sze (Bell Telephone Labs, Taiwan Chiao-Tung University), **Physics of Semiconductor Devices**, John Wiley & Sons, New York, 1st ed. 882pp, chapter 10 (1969), and 2nd ed. 868pp, chapter 8 (1981). This is the standard textbook on solid state devices since its first edition but has no circuits. It was written by an industrial researcher rather than a college teacher. Thus the treatment is compendious rather than tutorial and difficult for the beginners due to lack of physical explanation. It contains the state-of-the-art materials but the updates are infrequent to keep up with the rapid advances.

[699.3] William M. Penney and Lillian Lau (and American Microsystem Inc. technical staff), **MOS Integrated Circuits**, Van Nostrand Reinhold Co. New York, 1st ed, 474pp, (1972); reprinted by R.E.Krieger Publishing Co. Huntingon, NY (1979). This qualitative description of MOS device and circuits was written by the staff of the first MOS spring-off from Fairchild Semiconductor Corporation which started the Silicon Valley.

[699.4] Paul Richman (Standard Microsystems Corp.), **MOS Field-Effect Transistors and Integrated Circuits**, 259pp, John Wiley & Sons, New York (1973). This was the first and still the only MOS book that gives diagrams of electric field and charge distribution along the channel and near the drain junction in the current saturation range and the drain current-voltage characteristics in the punch-through range where the drain junction space-charge layer has thickened to reach the source junction (sections 4.3.n).

[699.5]^{f} Authur B. Glaser (Bell Labs.) and Gerald E. Subak-Sharpe (CCNY), **Integrated Circuit Engineering, Design, Fabrication and Applications**, 811pp, Addison-Wesley Publishing Co. Reading, MA (1977). This the second of the two textbooks covering MOS devices (sections 3.3-3.6) and circuits (sections 6.10-6.17 and 14.6.n). The first is Crawford's [699.2].

[699.6] Richard S. Muller (UC Berkeley) and Theodore I. Kamins (Hewlett-Packard Co.), **Device Electronics for Integrated Circuits**, John Wiley & Sons, New York, 1st ed (1977), 523pp. 2nd ed (1986). This is the one of three Berkeley textbooks on integrated circuits. MOSTs are covered in chapters 9 and 10 but no circuits. The treatments are not sufficiently quantitative and detailed to allow readers to follow the logical bases and the justifications of the approximations.

[699.7]^{dc} David A Hodges and Horace G. Jackson, **Analysis and Design of Digital Integrated Circuits**, McGraw-Hill Book Co. NY, 1st ed 434pp (1983) and 2nd ed 463pp (1988). This the second Berkeley textbook covering MOS device (chapter 2) and circuits (chapters, 3, 8, and 9).

[699.8]^{d} Yannis P. Tsividis (Columbia U.), **Operation and Modeling of the MOS Transistor**, McGraw-Hill Book Co. NY, 505pp (1987). Gives the most extensive small-signal MOST analysis.

[699.9]^{dc} Niantsu Wang (IBM), **Digital MOS Integrated Circuits, Design for Applications**, Prentice Hall, Englewood Cliffs, NJ, 361pp, (1989). Depth of the MOS device physics is average and inadequate for a beginner, but its MOST modeling and transient analysis are the most

comprehensive and accurate to-date which well serve as logical extension of our analyses in this chapter. IBM and not IEEE notation convention is used.

[699.10]^c David A. Hodges, editor (UC Berkeley), **Semiconductor Memory**, IEEE Press (1972). Covers a collection of $2^5=32$ articles in $2^3=8$ parts on MOS technology, devices, memory cell structures, peripheral circuits, and data sheets, up to 1972, providing a handy reference of hard-to-get historical articles and out-of-print data sheets.

[699.11]^c Mohamed I. Elmasry (U.Waterloo), **Digital MOS Integrated Circuits**, IEEE Press (1981). This contains 73 selected articles. They serve as practical examples of application of the MOST theory presented this chapter. A follow-up volume has not been published.

[699.12]^c Paul R. Gray and Robert G. Meyer (UC Berkeley), **Analysis and Design of Analog Integrated Circuits**, John Wiley & Sons, NY, 1st ed 681pp (1977) 2nd ed 771pp (1984). This Berkeley textbook is the companion to Hodges' digital circuit textbook [699.10]. The small-signal (section 1.8) and large-signal (section 1.7) MOST models show how device physics is used to carry out circuit analysis. However, the book focuses on bipolar circuit applications and leaves the following reprint volume [699.13] to cover MOS circuit applications.

[699.13]^c Paul R. Gray, David A. Hodges, and Robert W. Brodersen (UC Berkeley), **Analog MOS Integrated Circuits**, IEEE Press (1980). Paul R. Gray (UC Berkeley), Bruce A. Wooley (Stanford), and Robert W. Brodersen (UC Berkeley) **Analog MOS Integrated Circuits, II** (1988). These two reprint volumes give the most complete collection of historical articles on analog MOS integrated circuits. They supplement the preceding textbook by the same senior author.

[699.14]^c Roubik Gregorian (Sierra Semiconductor) and Gabor C. Temes (UCLA), **Analog MOS Integrated Circuits for Signal Processing**, 598pp, John Wiley & Sons (1988). On MOS filters.

[699.15]^c Lance A. Glasser (MIT) and Daniel W. Dobberpuhl (DEC), **The Design and Analysis of VLSI Circuits**, 473pp, Addison-Wesley Publishing Co., Reading MA (1985). The book author (Dobberpuhl) led the design of the MOS CPU processor chips for the first and later generations of VAX and MicroVAX computers manufactured by the Digital Equipment Corporation.

[699.16]^c Kit-man Cham, Soo-Young Oh, Daeje Chin, and John L. Moll (Hewlett-Packard Co.), **Computer-Aided Design and VLSI Device Development**, 314pp, Kluwer Academic Publishers, Boston (1986,1990). Describes the popular computer programs for MOS process and device simulation. Connects theory to numerical simulation methodology used in by manufacturing engineers.

[699.17]^c John Y. Chen (Santa Clara U., Xerox, Boeing), **CMOS Devices and Technology for VLSI**, 348pp, Prentice Hall, Englewood Cliffs, NJ (1990). The latest book solely on CMOS.

699 PROBLEMS

P641.1 By inspection without doing any algebra, write down the d.c. MOSFET equations for the p-channel device which are the pMOST counter part to (643.1) and (643.2) of nMOST.

P641.2 Give the detailed derivation of the pMOST d.c. equations using the procedure in the text that gave nMOST d.c. equations listed in (643.1) and (643.2).

P641.3 Use simple classical particle kinetics to show that the transverse electric field in the MOST is not a fundamental cause of mobility and drift velocity variation with the electric fields.

P651.1 Find maximum oscillation frequency formulae of a MOST using the one-lump hybrid-pi approximate small-signal equivalent circuit shown in Fig.651.2. Compare this frequency with the transit-time, transconductance, and gain-bandwidth cutoff frequencies described in section 652.

P654.1 Repeat P651.1 using the exact first-order small-signal pi model of Fig.654.1.

P660.1 The transconductance delay time (Fig.662.9), the transconductance and conductance cutoff or characteristic frequencies (Figs.664.2 to 664.4), and the normalized channel transit time (Fig.661.2) are all given in normalized units along the ordinate (y-axis). They are not dimensionless variables frequently used by theorists. Discuss the basic device physics underlying the reason on why they are not presented in dimensionless normalized unit but must use the normalization, L^2/μ_n, which has the dimension of Volt-second instead of second. The two-part answer is obvious if you examine the equations in the text that give the t's and ω's.

P660.2 In some parts of the text, C_L is used. In other parts of the text, C_D is used. Give one or more practical examples illustrating this multiple usage.

P660.3 The $i_D(t)$ transient shown in Fig.663.4(d) is incorrect. Without using any equations and use only the locus in Fig.663.4(b), give the obvious reason why $i_D(t)$ is incorrect. [Hint: The initial slope and the entire curvature are obviously wrong by examining the i_D-v_D locus in Fig.663.4(b).] Draw the correct $i_D(t)$ vs t using the $i_D(t)$ equation.

P660.4 If the drain and gate supply voltages are dropped to 0.6V and the oxide reduced to 40A as indicated by a recent (1989-1990) IBM 0.1-micron Si nMOST at 77K, compute the device parameters required in Table 662.1. Discuss the potential bit errors due to noise.

P660.5 Using the cross-sectional view of the MOSFET shown in Fig.620.1, illustrate the geometrical location and physical origin of the capacitances, resistances or conductances, and inductance of the small-signal equivalent circuit given in Fig.664.1. Reduce the distributed transmission line to the single lump representation given by this figure.

P660.6 Refer to Fig.661.1 and the nMOST given in section 654 with the following input data: $V_{DD}=+5.0V$, $V_{GT}=+2.0V$, $V_{GG}=+3.0V$ (from a d.c. supply not shown in Fig.661.1), $Z=20.0\mu m$, $L=5.0\mu m$, and $W/L\equiv Z/L=4.0$. Measured V_D is $+3.0V$. **(a)** Find the value of the load resistance, R=? ohms. **(b)** Compute the values at this bias point of the circuit elements of the small-signal hybrid-pi equivalent circuit of Fig.661.2 from the charge-control model presented in sections 654 and 661. **(c)** Repeat (b) and compute the additional parameters from the true small-signal admittance model given in section 664 and illustrated in the complete first-order hybrid-pi admittance equivalent circuit model shown in Fig.664.1. Tabulate your results from the charge control model obtained in (b) and the complete first-order model obtained in (c) in a table with two columns which contain g_{ms}, g_{mo}, g_{gd}, C_{gs}, C_{gd}, L_{ds}, $\omega_{gm}/2\pi$, $\omega_{gd}/2\pi$, $\omega_{ds}/2\pi$ and $\omega_{gs}/2\pi$. **(d)** Assuming that there is no capacitance loading, i.e. $C_d'=0$ in Fig.661.1, and using the charge-control small-signal model given in Fig.661.2, derive the formulae and/or sketch the drain voltage waveform, $v_D(t)$, when a step voltage of 5mV is superimposed onto the d.c. gate voltage, i.e. $v_G(t)=+3.0V+5.0u(t)$ where u(t) is the unit step function defined by $u(t\le 0)=0$ and $u(t>0)=1$.

P660.7 **(a)** Compute the intrinsic gate delay of the nMOST given in the above problem and also in section 661 at the biasing condition given above. **(b)** Compute the intrinsic power-delay product. **(c)** Does the load resistor increase the delay time computed here? Why?

P660.8 Use the nMOST in problem 660.6 but take V_G-$V_{GT}=V_{SS}=5.0V$ and let $C_L=1.0pF$. Compute and label the scales of all the capacitance charging and discharging curves given in Figs.663.4 and 663.5. {Note: This is a short problem since only the scale factors need to be computed to add the appropriate units to the axes of these figures, such as V, mA or pA, and ps or ns. Use the most convenient unit so that the axes labels are easy to read, i.e. 0, 1, 2, 3,... or 0, 0.1, 0.2, ...}

P660.9 Show that if the nMOST in problem P660.6, which has a large threshold voltage ($V_{GT}=+2.0V$), is used to charge up a capacitance C_L in a system with a maximum of 5V d.c., then C_L in Fig.663.3(a) cannot be fully charged up to 5V. What is the final steady-state voltage on this capacitor if $V_{SS}=+5.0V$, $V_{GT}=+2.0V$, and $v_G(t)=5u(t)V$. Give your derivation and give especially your reasoning in order to receive credit. {Hint: The reason is very simple and involves the MOST device physics on current saturation.}

P660.10 Suppose that the total effective load capacitance of a DRAM bit line is $C_L=10pF$ and is charged up to +5.0V. It is to be discharged by the nMOST given in problem P660.6 in the circuit shown in Fig.663.4(a) with a gate voltage step of 5V. The threshold voltage of the nMOST, V_{GT}, is +2.0V. Calculate and sketch the voltage and current waveforms. {Hint: Again the solution is very simple and takes little time to complete. Part of the answer is already given in Fig.663.4. Only the scales need to be relabeled with appropriate scale factors, a procedure already used in P660.9.}

P660.11 The dimension or unit of the MOST oxide capacitance C_o in the text is Farad/cm^2. However, in the complete one-lump small-signal model described in section 664, C_O is the total capacitance, namely, the areal capacitance times the gate or channel area, WxL. Verify this difference via dimension analysis of a selected equation (such as g_{mo} and g_{m}) from the charge control model in section 661 and from the true small-signal model in section 664.

P660.12 If V_G-V_{GT}=2V and V_D=5V are substituted into the grounded source MOST equation, (653.3), the current I_D is negative. Why is this incorrect and what is the basic physics?

P660.13 Why are the transconductance, g_m, given in (653.7) and (662.1) different?

P660.14 Can the transconductance in (653.7), (662.1) and (662.2) be used at any d.c. drain voltage V_D regardless of what the d.c. gate voltage is? Why?

P670.1 Refer to Fig.670.2, show that the MOSFET symbol is incorrect and misleading when there are three arrows, one each for the drain, source and substrate contact. Realize the physical structure that can be represented by such a three-arrow symbol.

P671.1 Describe the sequence of events that occur during the write-1 operation in a DRAM cell if the charge storage capacitance is already in the 1 state. Let 1 = +5V.

P671.2 Describe the sequence of events that occur during the write-0 operation in a DRAM cell if (a) C_S is already at 0-state and (b) C_S is at 1-state or +5V.

P671.3 Is there a redundancy of using the terms write-1, write-0, erase-1, and erase-0 in describing the DRAM cell operation? Which of these gives a unique set. Give an explicit example of voltage on C_S to illustrate such a unique set.

P671.4 Why is a restore operation not needed during a force read operation in a DRAM?

P671.5 Describe the sequence of events during a force-read operation when the charge stored on C_S is 0V. Following the description given in the text, give reasons as well as numerical estimates of time constants and initial, intermediate and final voltages.

P671.6 Instead of precharging the bit line to ½V_{SS} and using a dummy cell and a regenerative latch-up differential amplifier, precharge the bit line to V_{SS} without using a differential amplifier. Describe the sequence of events during a read operation of (a) a 0 on C_S and (b) a 1 on C_S.

P671.7 Calculate the refresh time interval if the Si MOS DRAM chip is operating at -55C assuming the same 0.1% maximum charge fluctuation on the storage capacitance C_S. What is the refresh time interval if it is to operate at 85C with not more than 0.1% charge fluctuation?

P671.8 Suppose that when charge fluctuations on Si DRAM cells reaches 50%, the memory bit error rate (still may be only 1% due to on chip and on board error correction circuits) will become so large as to cause a personal computer to crash. At what temperature will this occur? Like the text, use the SNS mechanism (i.e. n_i) to compute this maximum temperature.

P672.1 Derive the equations of output voltage vs input voltage, V_O vs V_I, and channel current vs input voltage, I_{CH} vs V_I, of the enhancement load NMOS inverter circuit (EENMOS) of Fig.672.22(c). Verify the results labeled on Fig.672.22(d) and (e). Use the procedure for the RENMOS and DENMOS inverters in the text and the loadline diagram of Fig.672.22(c).

P672.2 Draw to scale the channel current vs input voltage curves, I_{CH} vs V_I, of the three inverter circuits, RENMOS, DENMOS and EENMOS, using the equations derived in the text and in Problem 672.1. Label the important points and loci of these curves in a way similar to the labels for the V_O-V_I curves given in Fig.672.22(d). Use the sketched results of I_{CH}-V_I given in Fig.672.22(e) as a guide to label the curves.

P672.3 In the standby power dissipation estimate for the DENMOS inverter described in the text, it was implied that the voltage and current waveforms across the two nMOSTs follows instantaneously the dc current-voltage equation. Obtain the analytical expression for the dc standby power dissipation for all three inverters.

P672.4 Compute and draw the voltage and current transfer characteristics of the DENMOS, and EENMOS inverter circuits using the following data. $V_{DD}=+5V$. Input nMOST, $K_I=10mA/V^2$, $V_{TI}=+1V$. Load nMOST (enhancement-mode), $K_L=1mA/V^2$ and $V_{TL}=+1V$. Load nMOST (depletion-mode), $K_L=1mA/V^2$, $V_{TL}=-1V$. Plot the results on a linear coordinate paper with the following scales: 1V/inch and appropriate current scale per inch so that your results will display the characteristics well. What is the power dissipation during one switching transient? Use Figs.672.22(d) and (e) as guides so that you need not compute too many points.

P672.5 Compute and draw the voltage and current transfer characteristics of the CMOS inverter circuit using the following data. $V_{SS}=+5V$. $K_I=K_L=1mA/V^2$, $V_{TI}=-V_{TL}=1V$. Plot your results on linear graph paper. What is the power dissipation during one switching transient? Use Figs.672.23(d) and (e) as guides so that you need not compute too many points.

P672.6 In the text, we have not presented a transient analysis of the waveforms of the MOS inverter circuits. Give such an analysis for the CMOS inverter whose delay arises from the capacitance load and the intrinsic delay is unimportant. Assume a long-duration square wave for the input voltage. Repeat this problem for the more difficult but realistic case in which the input gate voltage is the output from a preceding CMOS driven by a square wave.

P673.1 Describe the latch condition of the 4-T store of a SRAM made of four PMOSTs instead of the four NMOSTs shown in Figs.673.1(a) and (b).

P673.2 Describe the read and write operations of a 6-T SRAM cell made of six pMOSTs instead of the six nMOSTs shown in Fig.673.1(c).

P673.3 Describe the latch condition and the read and write operations of the 6-T CMOS SRAM cell shown in Fig.673.2(d). Estimate the power dissipation of this CMOS RAM at a clock frequency of 100MHz for a 1-μm technology (gate, drain, source all have 1-μm_2 area) and an oxide thickness of 100A. Assuming negligible parasitics, what is the maximum clock frequency if there were no bit and word line capacitance delays?

P673.4 While looking through the latest issues (1990-2) of the Proceedings of the IEDM and ISSCC, one notices that many of the developmental SRAM chips use the RNMOS inverter. Give two or more reasons why this is preferred over the other three forms of SRAM (EE-NMOS, DE-NMOS, and CMOS). Does RNMOS have any drawbacks compared with one or all of the other three and why not during high speed operation? Use formulas and sketches. (Hints: PolySi R-load simpler topology, higher density, easy sheet resistance control, acceptable low standby power, comparable operation power at high clock frequency, and etc.)

P674.1 Explain how a Si UV-EPROM using n-channel could work and show why the p-channel proposed by Frohman-Bentchkowski is better. (Hint: SiO_2/Si electron barrier height is lower than hole.)

P674.2 Draw the energy band diagrams of the p-channel floating-gate UV-EPROM shown in Fig.674.1 during write, erase, and read.

P674.3 Draw the energy band diagrams of the n-channel flash EPROM shown in Fig.674.2 during write, erase, and read.

P674.4 Explain how an p-channel flash EPROM would work. Is the p-channel worse or better than the n-channel asides from speed? Estimate the write and erase times (using the Fowler-Nordheim tunneling formulae for a triangular barrier given in chapter 3) and give the oxide thicknesses and programming voltage for the same write and erase times to make the p-channel as fast as the n-channel.

P674.5 Let the texturized Si floating gate be modeled by semi-spherical hills. Calculate the radius of the hemisphere that would increase the write/erase time by 100 over the flat poly-Si/SiO_2 interface. Use the Fowler-Nordheim tunneling rate formulae. (Hint: Calculate the electric field at the tip of the hemisphere relative to the electric field at the plane. Your answer has two parameters, the average thickness of the tunnel oxide and the radius of the hemispherical Si intrusion into the oxide.)

P674.6 Explain the read, write, erase operations of the FMOSC memory cell shown in Fig.674.4(b).

P674.7 Is an initial charge up operation of the FMOSC memory cells necessary immediately after the d.c. power is turned on before it can be used for R/W operations.

P674.8 Explain the DRAM emulation mode of operation of the FMOSC memory cell.

P681.1 In the analyses on the effects of bulk charge, the threshold voltage equation given by (681.3) was given without a detailed derivation. This differs from the elementary derivation that gave (642.3) by the term $2V_F$. For the case of no voltage applied to the MOST, the result is identical to that of MOSC in chapter 3 such as (411.6) or (412.15) which used the exact surface field given by (412.6): $V_G = V_{FB} + V_S + (-sgU_S)\surd(2kT\varepsilon_s P_{EFF})$ where $P_{EFF} = \{(P_S-P_B) + U_S P_B\} + \{(N_S-N_B) - U_S N_B\}$, $U_S = qV_{I0}/kT$, $P_S = P_B \exp(-U_S)$, $N_S = N_B \exp(-U_S)$. (a) Verify this result at equilibrium and find the approximate solution when U_S is very large. Use the energy band diagram given in Fig.681.1(b) to do the following extensions. (b) Extend to the nonequilibrium case where V_S and V_X are no longer zero. (c) Extend to the general case where V_D is also applied.

P681.2 The bulk-charge or body effect on the MOST characteristics described in the text has two origins: (i) ionized doping impurity atoms in the Si surface space-charge layer which includes the thin surface channel and (ii) applied reverse bias voltage to the substrate. How is the model modified if the impurity concentration is zero, i.e. nMOST is made on pure Si with $N_{AA}=0$ and $N_{DD}=0$ except at the n+ drain and source? {Hint: Body effect (i) still exists when $V_X=0$ but the strong inversion model and the depletion approximation can no longer be used.}

P683.1 Use the parallel-plate capacitance formulae to compute the charge induced on the gate $(x=0)$ and on the Si surface $(x=x_o)$ by a sheet of oxide charge $\rho_{OT}(x)\Delta x$. Integrate this over the thickness of the oxide to give the total charge induced on the Si surface by a spatially distributed charged oxide traps, Q_{OT}, which was cited in (683.1).

P683.2 There is a sheet of charged oxide traps of $1.0\text{x}10^{11}q/\text{cm}^2$ located at the middle of a 100A oxide of a MOST. Draw the energy band diagram to scale. What is the gate voltage shift?

P683.3 A sheet of charged oxide traps of $1.0\text{x}10^{11}q/\text{cm}^2$ is migrating from the mid-oxide position towards the SiO_2/Si interface due to a high gate voltage applied to a 100A oxide. Draw the energy band diagram when the charge sheet reaches the SiO_2/Si interface. What is the gate voltage shift?

P683.4 Verify the arithmetic of the delta function approximation used to give an analytical formulae relating the gate voltage shift to the presence of interface traps, (683.9).

P683.5 Obtain the analytical correction term when the gate voltage shift of the subthreshold current is used to compute the interface trap density using (683.9). Assume that the substrate bias is not large so that an appreciable correction due to the bulk charge contribution of the threshold voltage must be made. Obtain a numerical estimate of the sensitivity or lower limit on the density of state in energy of the interface traps, D_{IT}, that can be detected if the drain current noise is 10%. Assume $V_{SX}=5$V, $N_{AA}=10^{10}\text{cm}_{-3}$, $n_i=10^{10}\text{cm}^{-3}$, and $x_o=100$ angstroms.

P684.1 Find the drain voltage dependence of the drain current in the medium longitudinal electric field range when the mobility is decreasing as $1/[1+E/E_{CL})^2]$.

P684.2 Find the effects of a large longitudinal electric field on the I_D-V_D characteristics when the bulk charge cannot be assumed constant.

P684.3 Find the effects of a large transverse electric field on the I_D-V_G characteristics when the bulk charge cannot be assumed constant. (i) Consider only increased phonon scattering due to 2-d phonons. (ii) Consider only interfacial surface roughness scattering.

P684.4 In general, the increased phonon scattering from 2-d phonons and surface roughness scattering operate simultaneously in a MOST. What is the numerical reason that these two can be analyzed individually and the results can be added using the Matthiessen rule?

P684.5 Derive the general analytical formulae of $I_D(V_D,V_G)$ when both 2-d phonon and surface roughness scatterings must be included. Hint: $I_D C_o L/\mu_0 W = \int(x-|Q_B|)dx/(1+bx+ax^2)$ where the integration variable is $x=Q_N+|Q_B|$, and the parameters are $b=Q_{CT}^{-1}$, and $a=Q_{SR}^{-2}$. Verify the asymptotic expressions for (i) $Q_N+|Q_B| << Q_{CT} << Q_{SR}$, (ii) $Q_{CT} << Q_N+|Q_B| << Q_{SR}$, and (iii) $Q_{CT} << Q_{SR} << Q_N+Q_B$.

P684.6 In general, the longitudinal electric field dependence due to the hot electron effect and the transverse electric field dependence due to increasing confinement occur simultaneously. Thus, they must be included in the MOST equation simultaneously. Derive the analytical expression of the $I_D(V_D,V_G)$ equation that accounts for both of these effects. Use as many scattering mechanisms as possible and still get an analytical solution.

P684.7 Confinement will actually increase the electron mobility due to less scattering by the ionized impurities in surface space-charge layer because of increased separation. This is most evident in the subthreshold range and just after the start of strong inversion, $V_{I0}=2V_F$. Using simple electrostatics, find the V_G and V_X dependences of the correction term due to confinement and increasing distance between the electrons and the bulk scatterers.

P685.1 Compute the drain conductance per unit gate width at $V_{DS}=0$ and $V_{DS}=5V$ with V_G-$V_{GT}=3V$ and the threshold voltage change at $V_{DS}=0$ due to the short channel effect in an Si n-channel MOST with $x_o=100A$, $L=1\mu m$, $N_{AA}=10^{18}cm^{-3}$, $N_{DD}>>N_{AA}$. Repeat for $L=0.1\mu m$ or somewhat larger if you find it necessary to give the normal (pentode-like) MOSFET characteristics. Check your results using the approximate formulae if you have made a more exact analysis.

P685.2 Compute the narrow gate effect on the threshold voltage and the drain conductance of two nMOST's with geometrical gate width of $W_G=1.0$ and $0.1\mu m$ at $V_{DS}=0$ and their drain currents at $V_{DS}=0.5V_{GS}=2.5V$. Use the parameters of the above submicron transistor. Check your results using the approximate formulae if you have made a more exact analysis.

P685.3 Show that (685.6) can be derived by letting $V(y=0)=V_S$ in (685.9).

P685.4 Extend the analyses of the two short-channel effects to include the voltage dependent bulk charge. Give two numerical examples using the submicron transistors of problem P685.1.

P685.5 Extend the analyses of the three narrow-gate effects to include the voltage dependent bulk charge. Give two numerical examples using the submicron transistors of problem P685.

P685.6 Extend the analyses of the two short channel effects and three narrow gate effects to the threshold range by including the voltage dependent bulk charge and interface surface potential. How do short channel and narrow gate affect the subthreshold slope of the I_D-V_{GS} characteristics?

P685.7 Design an LDD to eliminate or minimize the short channel effect in a $L=1.0$ and $L=0.1\mu m$ n-channel Si MOSFET. Assume spatially constant N_{AA} in each region of the device.

P685.8 Invent a structure that can completely eliminate the narrow gate effects. This is known as the fully recessed stepped oxide (FRSOX) whose sidewall oxide is thicker than the gate oxide and they are connected by a step or an abrupt change of oxide thickness. Show how practical fabrication limitations will modify this stepped oxide structure and how the non-ideal geometry would re-introduce the narrow gate effects. [See reference cited in Fig.685.5.]

P690.1 Sketch the geometric shapes of the conduction channel and the space-charge layers for the three bias conditions like Figs.690.1 for a dual-gate depletion-type FET which has a chemically doped channel, a top MOS gate and a bottom n/p+ junction gate.

P690.2 For the dual-gate hybrid MOS-JG-FET above, give the reasons of the very significant delay (time constant of the order of one millisecond) to switch the transistor from a depletion-mode bias condition to an enhancement-mode bias condition by a voltage step applied to the MOS gate. This is the low-frequency filtering mechanism discovered by Goppal Reddi in 1964 at Fairchild.

P690.3 Using the simple analysis described in the text, give the equations and plots of the d.c. output (I_D-V_D) and transfer (I_D-V_G) characteristics of the point-contact FET shown in Fig.690.1(d). Cover both the forward and reverse bias ranges of the d.c. voltage applied to emitter-base or source-gate junction. You have re-invented a new transistor if you succeed.

Chapter 7

BIPOLAR JUNCTION TRANSISTOR AND OTHER BIPOLAR TRANSISTOR DEVICES

Commenced First.
Ascending Again!

700 INTRODUCTION 704
710 BACKGROUND AND HISTORY 704
720 FABRICATION OF A DOUBLE DIFFUSED SILICON BJT 709
730 D.C. CHARACTERISTICS OF IDEAL AND REAL BJT'S 712
- 731 Two-Diode D.C. Circuit Representation of BJT, 712
- 732 Data of BJT Characteristics, 718
- 733 Derivation of the D.C. Characteristics of p/n/p BJT, 723
  - The Shockley BJT Equations, 730
  - The SNS BJT Equations, 731
- 734 The Original and Extended Ebers-Moll Equations of BJT, 733
- 735 Two-Port Nonlinear D.C. Network Representations of BJT, 739
  - General Common-Base Two-port Network Equations, 740
  - General Common-Emitter Two-port Network Equations, 742
  - Circuit Models in the Four Operation Modes, 743
    - Forward Active Mode, 743
    - Reverse Active Mode, 748
    - Cutoff Mode, 748
    - Saturation Mode, 749
- 736 Lumped D.C. Models of Realistic Multi-Dimensional BJT, 754
  - Non-overlap Collector Diode, 754
  - Base Spreading Resistance, 754
- 737 Material and Structural Dependences of D.C. Two-Port Parameters of BJT, 758
  - Diffusion-Drift Transport Time in the Quasi-Neutral Base Layer, 759
  - Gummel Number in the Quasi-Neutral Base Layer, 759
  - Gummel Number in the Quasi-Neutral Emitter Layer, 763
  - Emitter Injection Efficiency Calculated from the Gummel Numbers, 764
- 738 Bias Dependences of the D.C. Parameters of BJT, 765
  - The Early Effects in BJT and Lowly-Doped Collector, 765
  - The SNS Effects in BJT (α and β Fall-Off at Low Current), 770
  - Alpha and Beta Fall-Off at High Current in BJT, 774
  - The Kirk effects in BJT, 780
- 739 Collector Multiplication and Negative Resistance, 783
  - Numerical Example, 788

740 SMALL-SIGNAL CHARACTERISTICS OF BJT **791**
- The Small-Signal Condition, 791
- Comparing the Charge-Control and Exact Small-Signal Analyses, 793
  - Exact Analysis for Small-Signal Equivalent Circuit Models, 793
  - Charge-Control Analysis for Small-Signal Equivalent Circuit Models, 796

741 Common-Base Small-Signal Tee (CBss-Tee) Models of BJT, 797
- Low-Frequency CBss-Tee Model of Intrinsic BJT, 799
- High-Frequency CBss-Tee Model of Intrinsic BJT, 804
- CBss-Tee Equivalent Circuit Model of Real BJT, 811

742 Maximum Frequency of Oscillation of BJT, 814
- The Gibbons Frequency, 814

743 Common-Emitter Small-Signal Hybrid-Pi (CEss-Hπ) Model of BJT, 817
- The Conductance Elements of the CEss-Hπ BJT Model, 819
- Summary of BJT CEss-Hπ Conductance Elements and Numerical Example, 822
- Intrinsic Charge Control Capacitances of CEss-Hπ BJT Model, 823
- Parasitics of the CEss-Hπ BJT Model, 823

744 Common-Emitter Current Gain, Cutoff Frequency and Bandwidth, 823

750 LARGE-SIGNAL SWITCHING CHARACTERISTICS OF BJT **829**

751 The Diffusion and Charge-Control Equations, 830
- Charge-Control Equations of a General Slab, 832
- Charge-Control Equations of the Entire BJT, 832
- Charge-Control Equations of the Quasi-Neutral Base Layer, 832
- Charge-Control Equations of the Space-Charge Layers, 833
  - Emitter-Base Space-Charge Layer, 834
  - Collector-Base Space-Charge Layer, 834
- Complete Base-Charge-Control Equations, 834
- The Base Transport Time Parameters, t_{BF} and t_{BR}, 836
- Relating the Charge-Control and Ebers-Moll Parameters, 837

752 Common-Base Large-Signal BJT Switching Transients, 838
- Common-Base Turn-On Transients in BJT, 840
  - Phase I CB Turn-On Transient - Charge-Control Solution, 843
  - Phase II (including Phase I) CB Turn-On Transient - Lump Model, 848
  - Phase II CB Turn-On Transient - Exact Solutions, 851
  - Phase II CB Turn-On Transient - Effects of Parasitics, 854
- Common-Base Turn-Off Transients in BJT, 856
  - Charge-Control Solutions of CB Turn-Off Transients, 857
  - Exact Solutions of CB Turn-Off Transients, 860

753 Common-Emitter Large-Signal BJT Switching Transients, 862
- Common-Emitter Turn-On Transient, 862
- Capacitance Speed-up of Common-Emitter Turn-On Transient, 865
- Common-Emitter Turn-Off Transient, 866

754 Comparison of CB and CE BJT Switching Transients, 867
- Short-Circuit or Voltage Source in CB or CE Configuration (S,CE,CB), 867
- Open-Circuit or Current Source in CB Configuration (O,CB), 867
- Open-Circuit or Current Source in CE Configuration (O,CE), 867

755 Speeding Up the BJT via Technology, 869
- Technology of Reducing Recombination Lifetime τ_B, 872
- Speeding-Up via Geometry and Resistivity Reduction, 876
- Performance Comparison of Four Micron-Submicron BJTs, 878

756 Propagation Delay in Ring Oscillator, 882

760 CIRCUIT APPLICATIONS OF BIPOLAR JUNCTION TRANSISTOR 885
- 761 The BJT Digital Inverters, 885
- 762 The Common-Emitter BJT Inverter, 887
  - Charging Up the Emitter-Base Space-Charge Layer Capacitance, 888
  - Charging Up the Quasi-Neutral Base in the Active Range, 889
  - Further Charging Up the Quasi-Neutral Base in the Saturation Range, 890
  - Discharging the Stored Base Charge in the Saturation Range, 892
  - Discharging the Stored Base Charge in the Active Region, 895
  - Discharging the Space-Charge Layer Capacitances, 895
  - Average Propagation Delay in CE BJT Inverter, 896
- 763 Speeding Up the CE BJT Inverter, 897
  - Large Turn-On and Turn-Off Overdrives, 897
  - Input Speed-Up Capacitor, 897
  - Schottky-Barrier Bethe Diode Clamp, 897
- 764 The Emitter-Coupled 2-BJT Inverter (ECL), 899
- 765 The CB-CE Transistor-Transistor Coupled 2-BJT Inverter (TTL), 902
- 766 The Bipolar-MOS Inverters (BiMOS,BiCMOS,CBiCMOS), 906
  - Shunt-BiMOS: Emitter-Base Junction Shunted by MOST, 909
    Collector Follower, 911
    Emitter Inverter, 912
  - Series-BiMOS: MOST in Series with Base Terminal, 915
  - BiCMOS and CBiCMOS Inverters, 917
  - The Optimum 6-Transistor CBiCMOS Inverter, 928

770 THE HETEROSTRUCTURE BIPOLAR JUNCTION TRANSISTORS (HBJTs or HBTs) 931
- 771 Historical Background, 931
- 772 Fabrication Methods of Ge_xSi_{1-x} HBJT, 936
- 773 Operation Principle of HBJTs, 937
- 774 Energy Bands and Phonon Spectra of Commensurate Layers, 945

780 THE FOUR-LAYER PNPN DEVICES 955
  - Shockley's "Hook" Collector Theory, 960
  - Shockley's Four-Layer Diode Venture, 961
  - Silicon Controlled Rectifier (SCR), 962
  - Shorted-Emitter or Shunted-Emitter SCR, 964
  - Junction-Gate SCR, 965
  - Remote-Gate SCR, 965
  - Bilateral SCR, 965
- 781 Four-Layer PNPN Diode Characteristics, 965
  - D.C. Voltage-Current Characteristics of 4-Layer Diode, 966
  - Physics of the 4-Layer Diode V-I Characteristics, 970
  - Switching Transient in 4-Layer Diode, 972
- 782 PNPN Triode (SCR) Characteristics, 974
  - Basic SCR, 975
  - Shorted-Emitter or Shunted-Emitter SCR, 977
  - Junction-Gate SCR, 978
  - Remote-Gate SCR, 978
  - Bilateral SCR, 978
- 783 MOS-SCR, 979
- 784 Latchup in CMOS, 980

799 BIBLIOGRAPHY AND PROBLEMS 981

700 INTRODUCTION

The bipolar junction transistor (BJT) is historically the first solid state amplifier and switch manufactured in large volumes. However, its operation principle is more complex than the field-effect transistor, which also delayed its invention after the field-effect transistor. Once invented theoretically by Shockley in 1949, volume production began soon afterwards and large numbers were used in telephone switching offices for about a decade until BJTs took over in the 1960's. The more complex operation principle makes it pedagogically desirable to delay its discussion after the field-effect transistor. This is not the traditional sequence: it has usually been discussed first in the introductory and advanced electron device courses in college because it was the first solid- state amplifier and switch ever used in many applications. This book reverses the sequence not only for pedagogical reasons but also the fact that Si MOSFET has been and is expected to be the dominant device in applications using single transistors or integrated circuit chips. BJT will be used only in very high speed circuits and will occupy only a niche market. The background and history of BJT are discussed in section 710. A sample fabrication flow chart is described in section 720. The d.c., small-signal, and switching characteristics are described in sections 73n, 74n and 75n. Basic BJT digital circuit building blocks are described in sections 76n. Recent developments of extremely high speed (<10ps) and high frequency (>100GHz, mm-wave) heterojunction bipolar transistors in Ge_xSi_{1-x} and III-V compound semiconductor films are described in sections 77n. The four-layer p/n/p/n regenerative bipolar transistor known as the silicon controlled rectifier is briefly described in section 739. Problems and bibliography are given in section 799.

710 BACKGROUND AND HISTORY

Although the rectification property of a metal/semiconductor contact has been known since the 1920's and used as a radio signal detector, extensive applications did not begin until World War II (1939-1945) when it was used as the microwave detector in radar. Single crystal p/n junction diodes were not available commercially until the late 1940's and early 1950's because single crystals were difficult to grow. In addition to detector applications, diodes are also used as rectifiers to convert a.c. to d.c. as well as in switching applications.

Two terminal device, such as the p/n and metal/semiconductor junction diodes is severely limited when used as a switch in circuit applications because it does not have a third terminal to open or close the switch. In addition, diode with only one p/n junction does not amplify electrical signals since there isn't an internal positive feedback mechanism to give gain.

At the end of World War II, a focused effort was started at the Bell Telephone Laboratories (BTL) with the sole purpose of inventing a multiterminal solid state device which can amplify electrical signal and operate as a switch. The first three-terminal device tried in the laboratory was a thin-film insulated field-effect transistor fabricated by Pearson and Shockley in 1947. It was composed of a thin evaporated silicon or germanium film on a glass substrate. The electrical conductivity of the Si or Ge film was modulated by a transverse electric field created by a voltage applied to a field plate in parallel with the film. However, this thin film FET did not give amplification and power gain because of the high concentration of surface states on the semiconductor surface. These surface states trapped the mobile charges and substantially reduced the magnitude of conductivity modulation. Power gain was then observed by Bardeen and Brattain in 1947 in another field-effect transistor structure on a germanium single crystal. It has a surface channel connected to a metal/Ge point contact. The conductance of the surface channel was modulated by a ring-shaped field-effect gate electrode surrounding the point contact. The evolution of the field-effect transistors was recently reviewed by this author [600.1] and discussed in sections 610 and 690.

A still different transistor structure was also tried by Bardeen and Brattain in 1947 in which two closely-spaced metal points were placed on an n-type Ge single crystal substrate or base. One point contact was positive or forward biased and the other negative or reverse biased. Substantial current, voltage and power amplifications were recorded. A cross-sectional view is given in Fig.710.1(a). This was the first successfully manufactured solid state amplifier in human history. It was publicly announced in 1948 and coined the transistor following a suggestion made by John R. Pierce, the director of research at the Bell Telephone Laboratories at that time who was the inventor-developer of a wide-band microwave vacuum electronic amplifier known as the traveling wave tube. It turned out that this first point-contact transistor of Bardeen and Brattain was at least partially still another field-effect transistor. It was a composite transistor with two conduction pathways: the minority carrier pathway through the base and the majority carrier pathway along the surface channel described in Fig.690.1(d). In this first point contact transistor, the minority carrier pathway through the bulk or base of the single crystal Ge had probably contributed only a small percentage of the total current. Instead, the surface channel path may have dominated the current making it a p-channel junction-gate field-effect transistor. Its main conduction path was the induced p-type surface inversion channel on the n-Ge base and its input terminal was the p/n gate junction of the induced surface p/n junction between the induced p-type surface channel layer and the n-type base or substrate. This was discussed in section 690 and given as an advanced FET problem, P690.3.

Although this historically-first transistor amplifier with power gain was not a bipolar transistor based entirely on minority carrier injection, it led to the successful experimental demonstration of the bipolar transistor action from minority carriers by Shive at BTL in 1948. Instead of the two closely-spaced electrodes on one surface in Bardeen-Brattain's point contact transistor [Fig.710.1(a)], Shive placed the two metal electrodes on the opposite surfaces of a thin Ge wafer [like Fig.710.1(d)]. This became known as the coaxial transistor. The coaxial geometry substantially increased the surface path length between the two electrodes. The long surface path essentially eliminated the surface channel conduction which had dominated the current in the original Bardeen-Brattain point-contact transistor. Shive's coaxial transistor was the acknowledged first unambiguous existence proof of a bipolar transistor based entirely on the minority carrier injection action proposed by W= Shockley (William Shockley).

The theory of minority carrier injection was formulated by Shockley on April 24, 1947 which led to the theoretical invention of the bipolar junction transistor by Shockley in his **classic** article published in 1949. [W. Shockley, "Theory of p/n junctions in semiconductors and p/n junction transistors," Bell System Tech. Journal 28(7), 436-489, July 1949.]

The first bipolar junction transistor, manufactured in volume for commercial applications, was the point-contact transistor on n-Ge shown in Fig.710.1(a). It was used by the American Telephone and Telegraph Company in their telephone switching offices. Later, the **alloy junction bipolar transistor**, shown in Fig.710.1(b), was volume produced using a fabrication procedure invented by Dunlap and Hall at the General Electric Research Laboratory in 1950. The alloy BJT had two impurity-doped metal dots alloyed into the two surfaces of a thin Ge wafer to form two alloy-diffused p/n junctions. The two p/n junctions were actually diffused junctions whose impurity diffusants came from the acceptor impurity, such as Indium, in the metal alloy which formed the emitter and collector electrodes.

The first n/p/n **grown junction bipolar transistor**, shown in Fig.700.1(c), was demonstrated in 1951. Its thin p-type Ge base layer was produced during crystal growth by changing the dopant impurity concentration thrice from n-type to p-type and back to n-type. In addition to these two manufacturing methods, Philco Corporation used the metal/semiconductor Schottky barrier junction for both the emitter and the collector by electroplating the metal onto the two opposite surfaces of a thin Ge wafer as indicated in Fig.700.1(d). This was known as the **surface barrier transistor**. Its geometry was similar to Shive's coaxial point-contact transistor and it perhaps even operated similarly. The base thickness in the center of the Ge wafer of the surface barrier was precisely controlled by two coaxial jets of electrolyte which etch two small coaxial circular craters, one on each of two Ge surfaces. The base layer thickness between the two circular surface craters was determined by monitoring the attenuation of light through the base layer which was guided by the two coaxial electrolytic jets. After the desired base thickness was reached, the voltage polarity of the electrolytic jets relative to the Ge base was changed so that the metals are plated out onto the surface of the two craters to form the metal emitter and collector electrodes. The first thin-base 100 MHz surface barrier transistor was manufactured by this method, commercially marketed, and used in 88-108 MHz FM radio receivers. However, the electrolytic etching and plating operations on one transistor at a time were not amenable to mass production. In addition, the electrolytically plated metal/semiconductor junctions were not stable and difficult to reproduce. Furthermore, the principle of operation was and still is uncertain because the current of the metal/Ge Schottky barrier emitter could not contain many minority carriers.

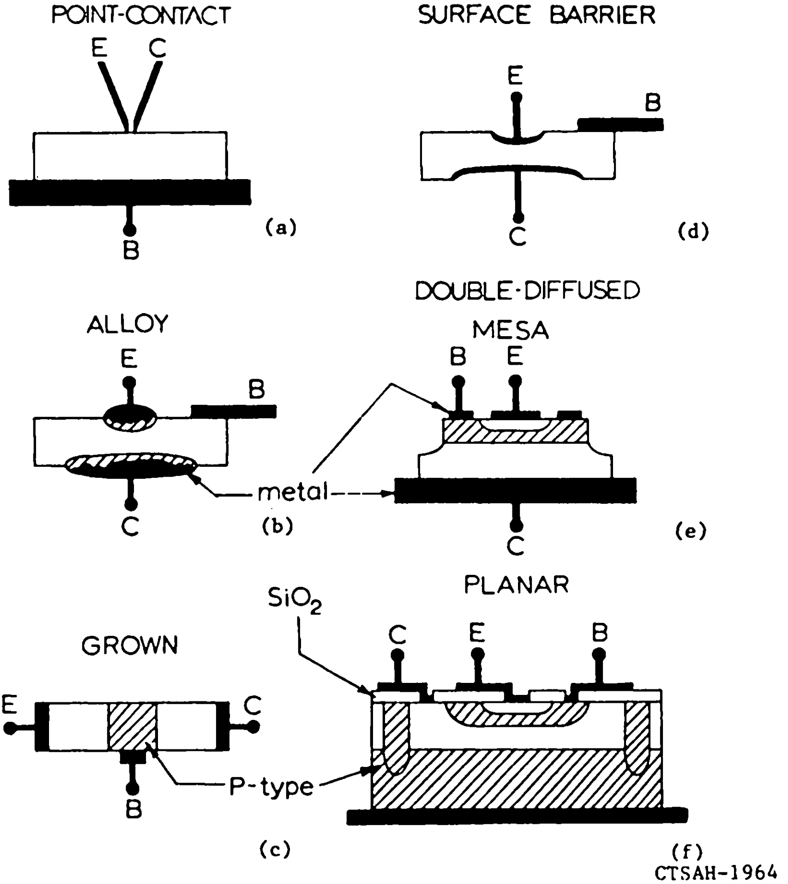


Fig.710.1 The historical development of the bipolar junction transistors. **(a)** the point-contact transistor on n-Ge. **(b)** The alloy-diffused junction transistor on n-Ge and p-Ge. **(c)** The grown junction transistor fabricated by alternating the dopant conductivity type during crystal growth. **(d)** The surface barrier transistor by jet electrolytic etching and plating. **(e)** The double-diffused Si mesa transistor. **(f)** The double-diffused oxide-passivated Si planar transistor.

The first three approaches, the point-contact, alloyed-diffused and grown junctions shown in Figs.710.1(a)-(c), could not give very thin base layer or very small emitter-collector separation. Consequently, these three transistors had low cut-off frequencies, in the audio range. The thinner base obtained by electrolytic etching in Fig.710.1(d) was not sufficiently thin, not consistently reproducible and not stable. Modern thin-base BJTs for high-frequency amplifiers were not built until 1956 when the development of solid state impurity diffusion technology began. Diffusion of impurity into a semiconductor from exposing its surface to a gaseous impurity source at a high temperature (about 600C in Ge and 1000C in Si) was first demonstrated by Scaff and Theuerer at Bell Telephone Laboratories (BTL) in 1951. The technique was then extensively developed for impurity diffusion into Si by Fuller and Ditzenberger at BTL in 1954 to fabricate device structures with multiple p/n junctions separated by very thin p-type and n-type layers. The slowness of impurity diffusion in crystalline semiconductors such as Si allowed highly accurate control of the thickness of the diffused layers. Base layer thickness of less than 10^{-4} cm, known as **submicron** today, was readily produced to give microwave and sub-nanosecond switching speed silicon bipolar junction transistors. Two diffused transistor structures were produced in succession. The diffused mesa transistor, shown in Fig.710.1(e), used a wax dot to mask the chemical etching of a mesa to isolate the collector junction. Oxide masking technique was employed to produce the modern diffused planar transistor structure shown in Fig.710.1(f). These have become commodity items in electronic hobby stores since the 1970's.

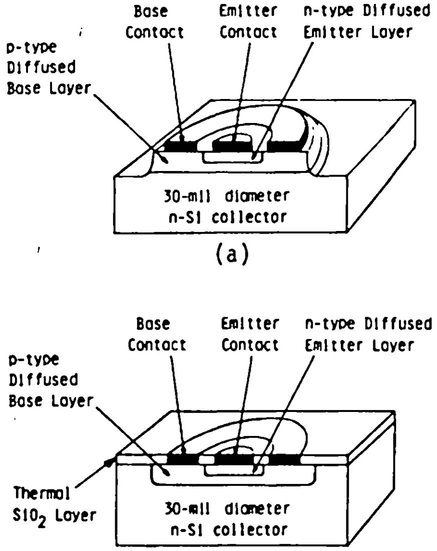
Using the solid state impurity diffusion technology and the oxide masking property against impurity penetration at the diffusion temperature (about 1000°C), silicon bipolar transistor integrated circuits containing a few to 100 logic gates on one silicon chip (small scale integrated circuits SSI) were then manufactured at the end of the 1960's and early 1970's. The transistor, diode, diffused resistor and interconnect patterns on the Si surface were produced by oxide masking and photolithography techniques. This circuit on one Si chip has been known as the **monolithic integrated circuit** and it was invented by Robert N. Noyce at the Fairchild Semiconductor Corporation in 1960 [600.1]. Noyce later co-founded Intel Corporation and was the first Chairman of SEMATECH when he died unexpectedly in the spring of 1990. The first monolithic Si logic gate chip was the resistor-transistor logic (RTL). The resistor was then replaced by a diffused Si p/n junction diode known as the diode-transistor logic (DTL). The diode was then replaced by a transistor giving the transistor-transistor logic (TTL) which was then clamped by a Schottky-barrier diode (S-TTL). TTL and S-TTL logic integrated circuits have dominated solid state electronics for two decades beginning in 1970 and is gradually being replaced by MOS and CMOS since 1980. Finally, the emitter-coupled logic (ECL) gate using silicon bipolar transistors was invented. Silicon ECL integrated circuit chip with 10,000 sub-nanosecond gates is the main component of the supercomputer CPU today.

The universally accepted acronym for the bipolar junction transistor is **BJT** emphasizing the need of two types of carriers, a majority and a minority carrier and their opposite polarities or electric charge (the electrons and holes). This is an unfortunate choice since some other supposedly not-bipolar transistors also operate on the same principle of minority carrier injection and also require the presence of majority carriers to control the minority carrier current. For example, the MOS field-effect transistor, when operated in the subthreshold range, utilizes a surface channel which is not inverted. Hence, the channel or drain terminal current is limited by the injection of the minority carriers from the source into the surface channel and by the diffusion of the injected minority carriers through the surface channel to reach the drain. This subthreshold diffusion current in MOSFET was analyzed in section 682. If the dominant mechanism is to be used to name a transistor, then the bipolar junction transistor is more appropriately called the minority-carrier injection transistor (MCIT or MINCIT) or a **bulk MCIT**. But, then the MOS transistor operating below surface channel inversion is also a MCIT or a **surface MCIT** but it is a MAJCIT (Majority carrier injection transistor) when the surface channel is inverted. Due to decades of usage, we shall follow the custom, BJT, in the following descriptions of the operation and properties of the bipolar junction transistor.

720 FABRICATION OF A DOUBLE DIFFUSED SILICON BJT

The first mass-produced double-diffused silicon BJT was the circular-shaped n/p/n Si mesa transistor whose cross-sectional view is shown in Fig.720.1(a). It was manufactured by the Fairchild Semiconductor Corporation in 1959 and had the part number 2N696. An early laboratory version was fabricated by Tanenbaum and Thomas in 1956 at Bell Telephone Laboratories (BTL) for internal use by the American Telephone and Telegraph Company. The word mesa, which describes the pedestal shape of the chemically etched transistor shown in Fig.720.1(a), was coined in 1954 by James M. Early who headed the transistor development group at BTL. Although the Si mesa BJT was the first mass-produced transistor of high performance (from thin base) and high temperature operation (from Si's larger energy gap than Ge's, see last part of section 242), recombination of electrons and holes at the surface states and in the surface channel on its exposed emitter-base junction surface had caused severe drop-off of the current gain at low currents. The bare collect-base junction surface also degraded the performance due to high surface leakage current and low junction breakdown voltage. Furthermore, the reliability is poor because the transistor characteristics were not stable due to the exposure of the unprotected junction surfaces to ambient which caused drift of the junction currents, increasing the leakage and decreasing the current gain. The reliability problems from the bare silicon surface was quickly overcome when a BTL group led by Atalla demonstrated stabilization in 1959 by protecting (passivating) the junction surface by thermally grown oxide during the diffusion of the base and emitter layers. This surface passivation technique was immediately applied to the Si BJT transistor by Hoerni in 1960 at the Fairchild Semiconductor Corporation who coined the oxide-passivated BJT the **planar transistor**. Subsequent 2N696's were all manufactured with oxide passivation, shown in Fig.720.1(b), although it was initially a mesa transistor, shown in Fig.720.1(a).

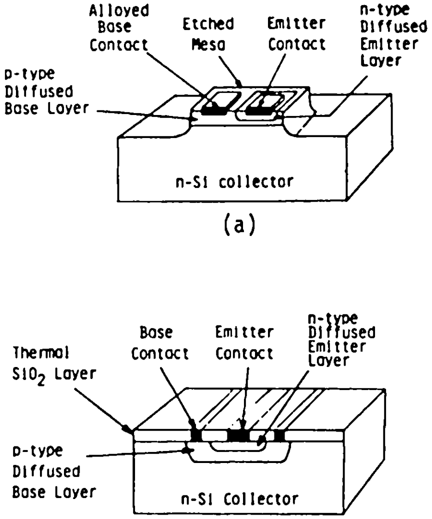
Typical high-speed or high-frequency Si planar transistors have a linear or stripe geometry instead of circular even during the mesa era, Fig.720.2(a), because the stripe geometry is easier to design and more efficient in using real-estate. The linear planar BJT with two base contact stripes is shown in Fig.720.2(b). Today, the stripe width has shrunk to less than 0.5 micron for extremely high speed (subnanosecond) integrated circuits engraved by electron-beam lithography.


Base Contact
Emitter Contact
n-type Diffused Emitter Layer
p-type Diffused Base Layer
30-mil diameter n-Si collector
Base Contact
Emitter Contact
n-type Diffused Emitter Layer
p-type Diffused Base Layer
Thermal SiO_2 Layer
30-mil diameter n-Si collector

(a)

(b)

Fig.720.1


Alloyed Base Contact
Etched Mesa
Emitter Contact
n-type Diffused Emitter Layer
p-type Diffused Base Layer
n-Si collector
Thermal SiO_2 Layer
Base Contact
Emitter Contact
n-type Diffused Emitter Layer
p-type Diffused Base Layer
n-Si Collector

(a)

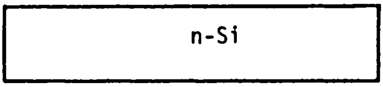
(b)

Fig.720.2

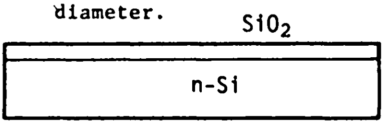
Fig.720.1 The cutaway views of the double-diffused Si n/p/n transistors 2N696 manufactured by Fairchild Semiconductor Corp. in 1960. **(a)** Mesa. **(b)** Planar.

Fig.720.2 Si n/p/p BJT using linear geometry to attain high speed in the 1960's. **(a)** 1-stripe mesa. **(b)** 2-stripe planar, the 2N709 @ 3ns designed for the first production supercomputer CDC-6600. (1-mil = 10^{-3} inch.)

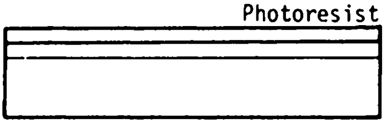
A typical fabrication sequence of a modern double diffused silicon planar n+/p/n BJT is shown in Figs.720.3(a) to (n). Some of the impurity predeposition or prediffusion steps have been replaced by ion implantation which gives a more precise control of the impurity concentration and junction depth and allows better control than diffusion especially at low concentrations. The fabrication steps are stated by the label next to each figure. A VLSI technology textbook can be consulted for the basic solid-state and liquid-state chemistries which explain the mechanics and controlling mechanisms of each of the processing steps in order to give reproducible structures for reliable, high-performance transistor characteristics.



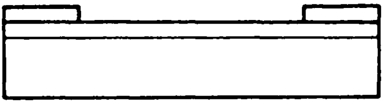
(a) n-Si slice. 2-3" (1959-61) diameter.



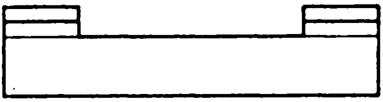
(b) SiO_2 thermally grown, 1000C.



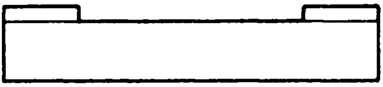
(c) Photoresist applied.



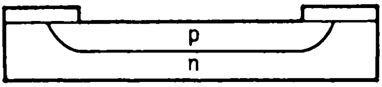
(d) Photoresist exposed and developed.



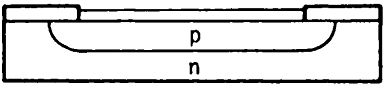
(e) Base diffusion window etched in SiO_2.



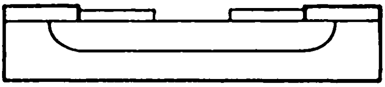
(f) Photoresist removed.



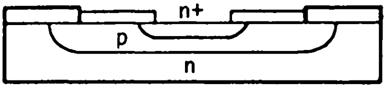
(g) p-type base impurity diffused.



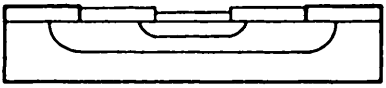
(h) Surface reoxidized.



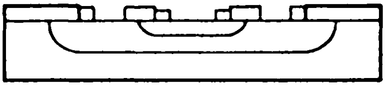
(i) Emitter window etched in SiO_2 by repeating steps (c) to (f).



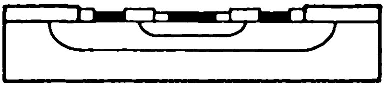
(j) n+ emitter impurity diffused.



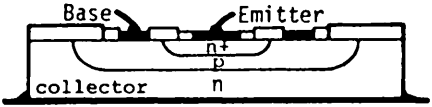
(k) Surface reoxidized.



(l) Base and emitter contact areas etched by repeating (c) to (f).



(m) Aluminum evaporated and contact area defined using (c) to (f).



(n) Slice diced into $50x50mil^2$ chips. Die attach on TO-18 header and wire bond with 1-mil gold wire.

Fig.720.3 (a) to (n) The cross-sectional view and the processing step at each manufacturing stage of an oxide-passivated, double-diffused, silicon planar transistor.

730 D.C. CHARACTERISTICS OF IDEAL AND REAL BJT'S

The d.c. characteristics and the circuit representations of the BJT are described and derived in the following sections, 73n. The ideal one-dimensional intrinsic transistor will be described and analyzed first, followed by an analysis of the two-dimensional effects using regional one-dimensional models. Effects of material properties, and d.c. biasing voltage and current levels on the characteristics will also be described and analyzed. The analyses are simple and physics-based to bring out the fundamental mechanisms that determine the d.c. electrical characteristics. Circuit representation will be used to bring out the basic physics.

731 Two-Diode D.C. Circuit Representation of BJT

We shall first give a qualitative description of the d.c. current-voltage characteristics of the BJT based on physical intuitions derived from what we have learned about p/n junction diodes with a minimum amount of mathematics. An amazingly concise and accurate model can be developed since a BJT consists of two closely spaced p/n junction diodes connected back-to-back to share the same n-type region. This proximity gives rise to the transistor action which is absent when the two diodes are far apart. A more detailed mathematical analysis for the ideal one-dimensional (1-d) BJT will then be given in the next section. The IEEE circuit and device convention of notations and symbols are used. The **independent** sources are enclosed by **circles**. The **dependent sources** are enclosed by **diamonds**. The nature of the source is indicated by the device symbol enclosed by the circle or the diamond. For example: the **arrow indicates a current source**; the +/- sign indicates a **voltage source**; and the rectifying p/n junction diode symbol, —|◀—, indicates a **nonlinear diode**.

Figures 731.1 to 731.4 show the cross-sectional view and equivalent circuit models of the p/n/p and the n/p/n BJTs in the **common-base** and **common-emitter circuit configurations**. The p/n/p transistor is now described in detail.

Figure 731.1(a) shows a thin cross-section cutting through the **two junctions** (emitter and collector) and **three layers** (emitter, base and collector) of the p/n/p BJT. It is not drawn to scale. The **collector** layer is the substrate or the body of the silicon wafer which is the thickest (500 micrometers or 25 mils). The **emitter** and **base** layers are thin (about 1 micrometer or less). The thin layers are produced by impurity diffusion, ion implantation and epitaxial growth techniques.

Figure 731.1(b) is the consensus **circuit symbol** of the p/n/p transistor. It gives also the positive reference direction of the three currents, I_E, I_B, and I_C. The positive reference direction is the direction of positive charge flowing into a terminal. In some transistor circuit textbooks, the convention for the collector current is opposite to ours in order to make all three currents positive when the transistor is operated as an amplifier.

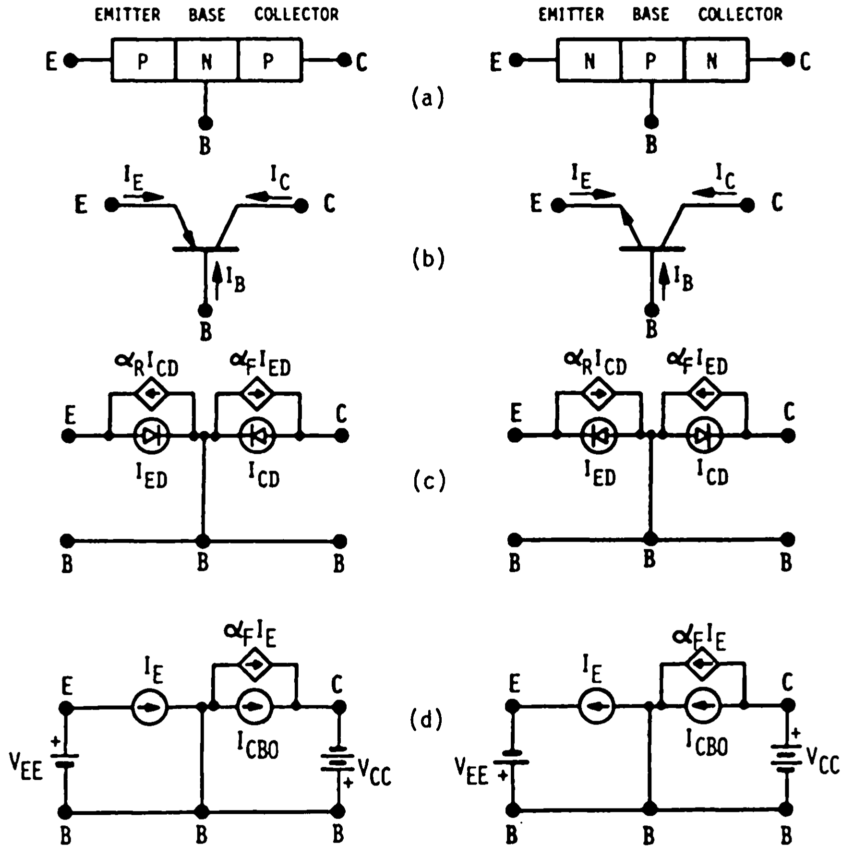

EMITTER BASE COLLECTOR
E
P
N
P
C
B
EMITTER BASE COLLECTOR
E
N
P
N
C
B
(a)
I_E
I_C
E
C
I_B
B
(b)
$\alpha_R I_{CD}$
$\alpha_F I_{ED}$
E
C
I_{ED}
I_{CD}
B
B
B
(c)
$\alpha_F I_E$
I_E
E
C
I_{CBO}
V_{EE}
V_{CC}
+
+
B
B
B
(d)

Fig.731.1 **Fig.731.2**

Figs.731.1 and 731.2 The p/n/p and n/p/n BJTs in the common-base circuit configuration. **(a)** A thin cross-sectional view. **(b)** The circuit symbol. **(c)** The general d.c. equivalent circuit. **(d)** The d.c. equivalent circuit in the forward active mode.

The term, **base**, comes from the 1948 point contact transistor invented by Bardeen and Brattain which was made of two metal wire whiskers pressured into contact with an n-type Germanium (base). One of the point contacts is positively biased relative to the base in order to emit holes into the n-Ge base, so named **emitter**. The other point contact is negatively biased relative to the n-Ge base in order to collect the holes injected by the emitter, so named **collector**. The current-voltage rectification characteristics of the two metal/n-Ge point-contact diodes suggested the circuit symbol shown in Fig.731.1(b). The arrow in the emitter terminal shows the direction of easy current flow. The arrow in the collector terminal is intentionally deleted for two reasons. (1) An arrow for easy current direction in the collector-base junction diode would be pointing in the direction opposite to the main current flowing through the collector-base junction which comes from the minority carriers injected by the emitter-base junction. (2) A collector diode arrow would make the transistor symbol symmetrical which obscures the fact that most practical transistors are asymmetrical:- different emitter and collector geometries (area and thickness) and material properties (dopant impurity concentrations). Even the base layer has an asymmetrical dopant impurity concentration profile. These asymmetries cause large differences in the **forward** and **reverse** electrical characteristics, such as the breakdown voltage and leakage current of the emitter-base and collector-base junctions, and the signal amplification or gain. Reverse (or backward) means that the input is applied to the collector and the output is taken from the emitter in the common-base configuration.

Figure 731.1(c) is the general d.c. equivalent circuit of the p/n/p transistor in the **common-base (CB) circuit** or **bias configuration**. The base terminal is the **common** terminal for applying or measuring the d.c. voltage source and time-varying signal at the emitter input and collector output terminals.

This d.c. equivalent circuit can be readily deduced by decomposing the collector and emitter **terminal** current into two components, one caused by the **diode** current flowing in the collector p/n junction and the other, the emitter junction. The current flowing into the collector terminal, I_C, consists of two components. (1) The first component of I_C is the diode current flowing through the collector-base p/n junction diode due to the voltage, V_{CB}, applied between the collector-base terminals, denoted by I_{CD} where subscript D stands for diode to distinguish it from the total collector terminal current I_C. I_{CD} consists of the Shockley ideal diode current due to minority carrier generation or recombination and diffusion in the quasi-neutral collector and base layers, and the SNS diode current due to electron-hole generation or recombination in the collector-base junction space-charge layer. (2) The second component of I_C is due to holes crossing the collector-base junction in the hard direction (from n-base to p-collector). These holes are injected from the p-type emitter into the n-type base layer by the voltage V_{EB} applied between the emitter and base terminals. (2) constitutes only a fraction, α_F, of the total emitter-base-junction diode current, I_{ED}. $1-\alpha_F$ accounts for the fraction of the injected holes that are lost by recombination with electrons in the emitter-base-junction space-charge layer and the quasi-neutral n-type base layer. $1-\alpha_F$ also accounts for the part of the emitter diode current due to electrons injected from the n-base into the quasi-neutral p-emitter layer towards the emitter terminal. Thus, the total collector terminal current is given by

$$I_C = I_{CD} - \alpha_F I_{ED}. \qquad (731.1)$$

Similarly,

$$I_E = I_{ED} - \alpha_R I_{CD}. \qquad (731.2)$$

These two equations give the equivalent circuit model shown in Fig.731.1(c).

The diode currents, I_{ED} and I_{CD}, are not exactly the sum of the ideal Shockley diode equation, (532.14), and the SNS diode equation, (535.4), since the Shockley diode equation (532.14) was derived for minority carrier currents flowing in the thick p-type and n-type quasi-neutral layers of a thick diode while the emitter-base diode of a BJT has a thin base layer and usually a thin emitter layer (in diffused or epitaxial BJTs) and the collector-base diode of a BJT has a thin base layer and sometimes a thin collector layer also (in BJT with a heavily-doped buried collector contact below the thin base layer).

Figure 731.1(d) is the approximate d.c. equivalent circuit of a p/n/p BJT in the CB bias configuration operating in the **forward active** bias **mode**. This is the amplification mode that gave the transistor action discovered by Bardeen and Brattain. The forward active mode is maintained by the polarity of the two d.c. bias voltages shown in this figure. The emitter-base junction is forward biased in order to inject the holes from the p-type emitter into the n-type base. The collector-base junction is reverse biased, in order to collect a major fraction of the holes injected by the emitter. Thus, the collector current consists of two parts from the same two origins which gave us the two current sources given by (731.2) and shown in the generalized d.c. equivalent circuit model, Fig.731.1(c). One is the leakage current due to the reverse-biased collector-base junction and the other is the current due to the emitter injected carriers passing through the collector-base junction. Thus, the collector current is given by

$$I_C = I_{CBO} - \alpha_F I_E. \qquad (731.3)$$

I_{CBO} is the **collector-base junction reverse leakage current** in the common-base configuration (subscripts C and B) with the other junction or terminal open-circuited (subscript O). Open circuit here means $I_E=0$ or the emitter terminal (input node) is floating. $\alpha_F I_E$ is the fraction of emitter current that reaches and is collected by the collector-base junction. The negative sign comes from our convention of pointing the positive current into the transistor collector terminal. α_F is the common-base transistor current gain in the forward direction known as the **forward alpha**. The term forward and subscript F are added in order to distinguish it from the **reverse alpha**, α_R, which is the current gain when the transistor in connected in the reverse or backward configuration, that is, the input signal is applied to the collector and the output is taken from the emitter. I_E consists of both the emitter diode current, I_{ED}, and the collector diode leakage current that reaches the emitter which is negligible. Thus, the equivalent circuit in the forward active mode shown in Fig.731.1(d) can be immediately drawn using (731.3).

In summary, the key principles and mechanisms in BJT operation and **transistor action**, as envisaged by Bardeen, Brattain and Shockley are: (1) the **injection** of minority carriers into a thin semiconductor base layer by a low input signal voltage applied across a forward biased emitter p/n junction and (2) the **collection** of most of the injected minority carriers by a reverse-biased n/p collector-base junction whose high electric field sweeps the injected carriers quickly through the junction space-charge layer. The key fundamental phenomenon is **minority carrier injection**, discovered and first analyzed mathematically by Shockley on April 24, 1947 [600.1], and it has been emphasized by the repeated use of the term **injection** in subsequent literature. For example, the p/n junction light emitting diode (LED) that lases is known as the injection diode laser.

Simple algebra will give the equivalent circuit model in the **common-emitter** (CE) bias configuration which has the base terminal as the input and collector terminal as the output. The CE configuration is the one most commonly used in both analog (small-signal linear) and digital (large-signal switching) circuit applications because of its high voltage-gain and large fan-out. Large fan-out means one transistor can drive several transistors without appreciable slow-down and losss of signal amplitude by multi-transistor loading. Circuit stage using BJT in the CE configuration can have a large fan-out owing to the high input impedance of the base terminal of the BJT compared with the emitter terminal in the CE configuration.

Figures 731.3(a)-(d) show the CE configuration of the p/n/p BJT. Figures (a), (b) and (c) are identical to those of the CB configuration, given in Figs.731.1(a)-(c), except for a 90 degree rotation and mirror reflection. The equivalent circuit of the CE forward active mode given in Fig.731.3(d) differs from that of the CB given in Fig.731.1(d). The difference is in the two current sources in the collector terminal branch since the input variable is the base current, I_B, in the CE configuration instead of I_E in the CB configuration. The two collector current sources in the CE configuration can be readily derived from the CE current equation, (731.3), using $I_E = -(I_B+I_C)$,

$$I_C = I_{CBO} - \alpha_F I_E = I_{CBO} - \alpha_F[-(I_B+I_C)]$$

giving

$$\begin{aligned} I_C &= I_{CBO}/(1-\alpha_F) + [\alpha_F/(1-\alpha_F)]I_B \\ &= (\beta_F+1)I_{CBO} + \beta_F I_B \\ &= I_{CEO} + \beta_F I_B \end{aligned} \tag{731.4}$$

where

$$I_{CEO} = (\beta_F+1)I_{CBO} \tag{731.4A}$$

and

$$\beta_F = \alpha_F/(1-\alpha_F). \tag{731.4B}$$

I_{CEO} is the collector leakage current in the CE configuration with the base input open-circuited or $I_B=0$. β_F is the d.c. forward current amplification factor in the CE configuration, also known as **forward** d.c. **beta** and denoted by h_{FE} in the hybrid two-port representation described in section 735.

The two current equations, (731.1) for I_C and (731.2) for I_E, give a concise and complete description of an intrinsic one-dimensional BJT. The detailed analysis of the nonlinear BJT d.c. current-voltage characteristics in the CB and CE configurations will be given in section 735 after we have derived the formulae and delineated the physics (the controlling transport mechanisms) of the diode currents, I_{ED} and I_{CD}, and the current gains, α_F and α_R in sections 733 and 734. First, the d.c. data will be illustrated and described in the next section, 732.

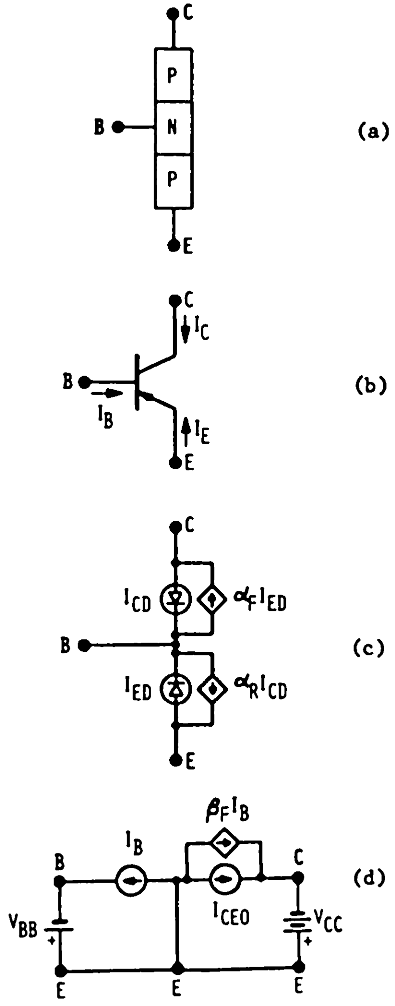

C
P
N
P
B
E
(a)
C
I_C
B
I_B
I_E
E
(b)
C
I_{CD}
$\alpha_F I_{ED}$
B
I_{ED}
$\alpha_R I_{CD}$
E
(c)
$\beta_F I_B$
I_B
B
C
(d)
I_{CEO}
V_{BB}
V_{CC}
E
E
E

Fig.731.3

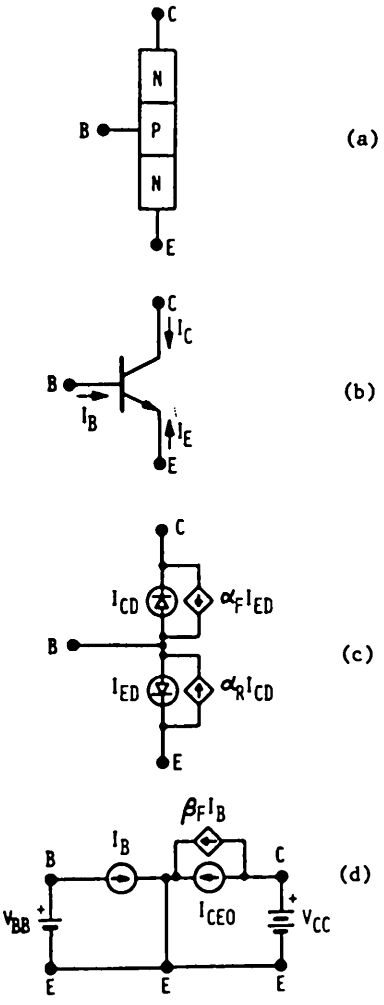

C
N
P
N
B
E
(a)
C
I_C
B
I_B
I_E
E
(b)
C
I_{CD}
$\alpha_F I_{ED}$
B
I_{ED}
$\alpha_R I_{CD}$
E
(c)
$\beta_F I_B$
I_B
B
C
(d)
I_{CEO}
V_{BB}
V_{CC}
E
E
E

Fig.731.4

Figs.731.3 and 731.4 The p/n/p and n/p/n BJTs in the common-emitter circuit configuration. **(a)** A thin cross-sectional view. **(b)** The circuit symbol. **(c)** The general d.c. equivalent circuit. **(d)** The d.c. equivalent circuit in the forward active mode.

732 Data of BJT Characteristics

Before going into the details of the physics and the elementary mathematics to derive the equations that give the characteristics of BJTs, we shall first describe the d.c. steady-state current-voltage characteristics of a typical n/p/n BJT transistor displayed by a transistor curve tracer. These include the two families of output characteristics (input current or voltage as a constant parameter), and the three sets of transfer characteristics (conductance-transfer or transconductance, current-transfer and voltage-transfer), for the forward (normal) and reverse (backward, abnormal, or inverse) circuit connections in the common-base and common-emitter configurations. This gives a total of twenty families [(2+3)x2x2=20] of d.c. curves for an n/p/n transistor. Twenty similar families of curves are obtained for a p/n/p transistor. In later sections, enlarged views are used to describe the physics and the analyses of the effects of material, geometry and bias voltage and current on these transistor characteristics.

The **output** current-voltage **characteristics** of an n/p/n BJT in the **common-base** connection are shown in Figs.732.1(a)-(a') and (b)-(b'). Figures (a) and (a') are the output characteristics, I_C vs V_{CB}, in the **forward** connection, defined by using the emitter-base terminals as the input and collector-base terminals as the output. Each curve in (a) has the emitter current kept constant while in (a') the emitter-base voltage is kept constant. Figures (b) and (b') are the output characteristics in the **reverse** connection, I_E vs V_{EB}, defined by using the collector-base terminals as the input and emitter-base terminals as the output.

Consider figure (a). It contains a family of I_C vs V_{CB} curves of an n/p/n BJT with the emitter current, I_E, kept constant for each curve. The four regions of transistor operation extend over three of the four quadrants of the I-V plane. The first quadrant $(x>0,y>0)$ is the **active region** where amplification and power gain can be obtained.

The **saturation region** is in the second quadrant $(x<0,y>0)$. In this region the transistor switch is closed or the transistor is turned on by applying a large input current. This moves the I_C-V_{CB} characteristics to follow the $V_{CB}<0$ nearly-vertical border line on the left. The collector-base junction is now forward biased ($V_{CB}<0$ for n/p/n) which gives a very low output resistance that essentially short circuits the output (collector terminal) to the common or ground (or base) terminal.

The third region is in the third quadrant $(x<0,y<0)$ and represented by a single line or locus called the **short-circuit line** or **locus** along which both the emitter and the collector junctions are forward-biased and have very low resistances. It is a continuation of the border line of the second region, the saturation region.

The fourth region is slightly above the positive x axis in the first quadrant $(x>0,0+>y\geq 0)$, and represented by a single line termed the **cut-off line**. Along this line, the transistor passes essentially no input and output currents, i.e., the transistor switch is open or the transistor is turned off.

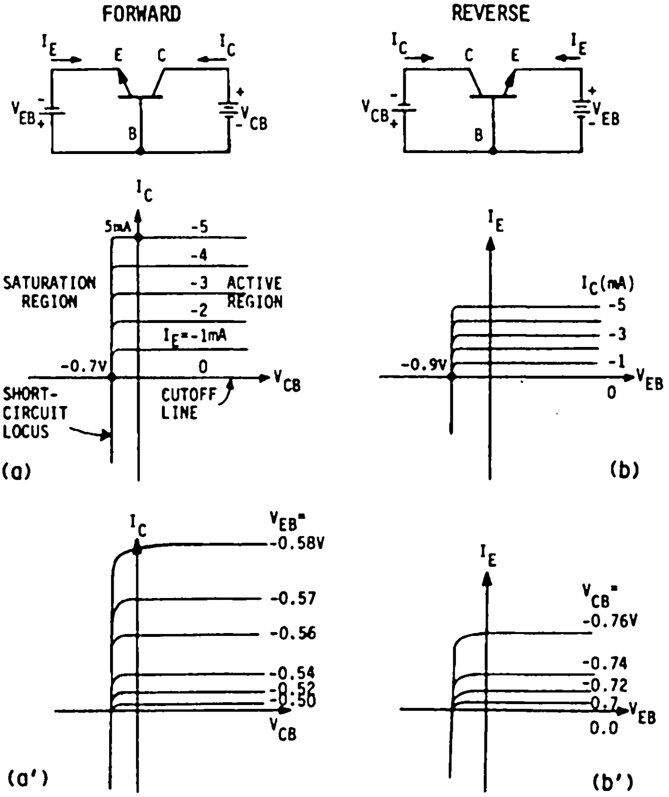


Fig.732.1 The common-base d.c. output characteristics of an Si n/p/n bipolar junction transistor. (a) The forward connection and (b) the reverse (backward) connection.

These four regions of operation are all labeled in Figs.732.1(a) for the forward connection just discussed. Similar labels can be added for the forward and reverse connections in Figs.732.1(a), (b) and (b'). Notice the following two features. (1) The collector current is nearly equal to the magnitude of the emitter current in the forward active region in figure (a). This is expected since the collector current comes from the electrons (minority carriers in the p-base layer) injected by the emitter-base junction, and most electrons reach and are collected by the collector-base junction space-charge layer. The nearly equal I_C and I_E reflects the low recombination loss of electrons with holes in the p-base layer, resulting in a nearly unity forward alpha: $\alpha_F = 1$. (2) The collector-base voltage is negative or the collector-base n/p junction is forward biased in the saturation region although the output conductance (dI_C/dV_{CB}) is still very small before the voltage drops to the short-circuit locus. The negative sign in I_E comes from our choice of the reference direction of current: positive is the direction of positive charge or positive current flowing into the transistor terminal from the external wire.

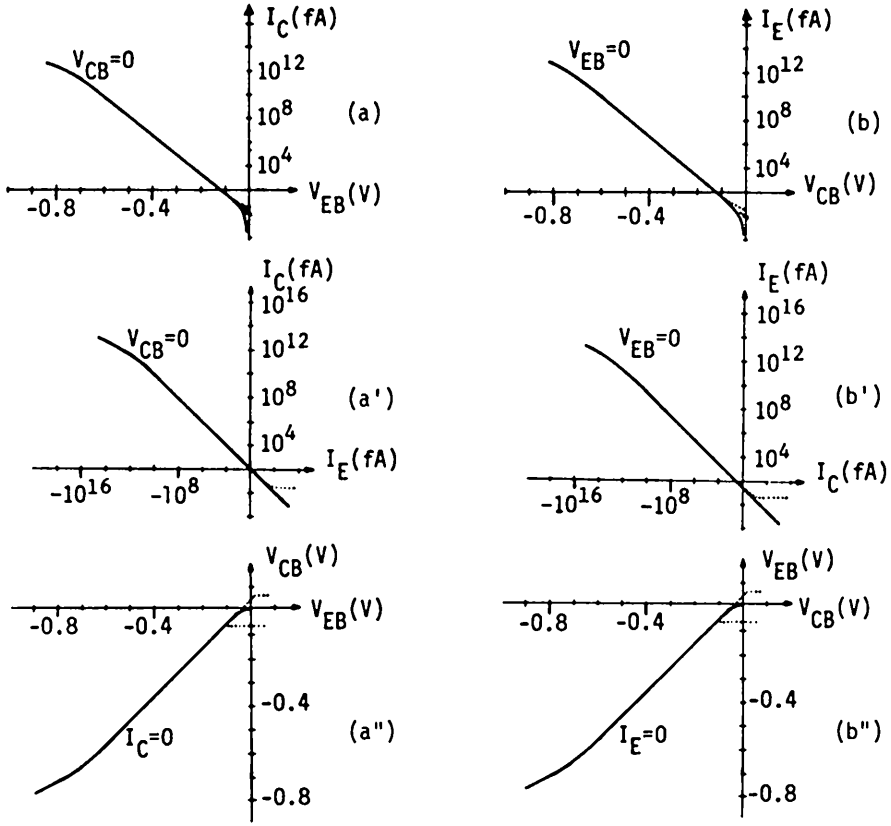


Fig.732.2 The common-base d.c. **transfer** characteristics of an Si n/p/n bipolar junction transistor. (a) conductance transfer or transconductance, I_C-V_{EB}, (a') current transfer, I_C-I_E, and (a") voltage transfer, V_{CB}-V_{EB}, characteristics in the forward configuration. (b), (b'), and (b") are in the reverse configuration. Dotted lines are those with a reverse-bias or -current parameter.

The d.c. **transfer** characteristics of the same Si n/p/n transistor are shown in Figs.732.2(a), (a'), (a"), (b), (b') and (b"). (a) is the **forward conductance-transfer** characteristics whose slope gives the transconductance of the BJT. (a') is the **forward current-transfer** characteristics whose slope gives the differential alpha, α_f, and whose ratio, I_C/I_E, gives the d.c. alpha, α_F, if $I_{CBO}=0$. (a") is the

forward voltage-transfer characteristics which gives the floating potential across the CB junction when a forward voltage is applied to the EB junction. The slope gives forward d.c. alpha, α_F. The corresponding three transfer characteristics in the reverse configuration are shown in figures (b), (b'), and (b").

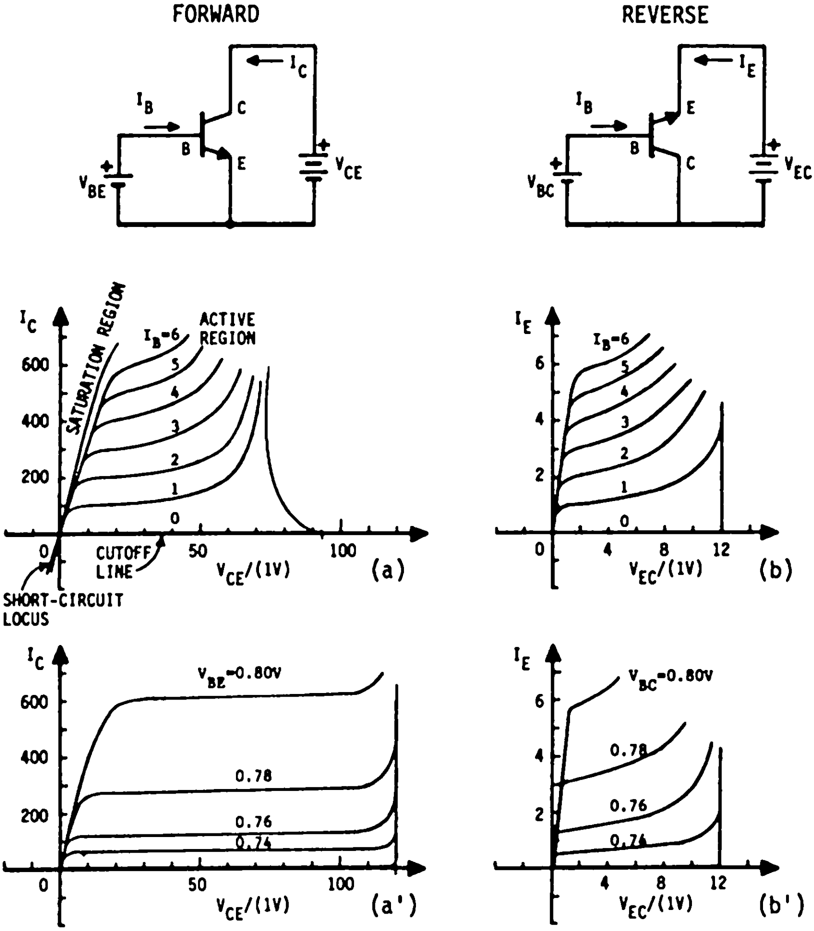


Fig.732.3 The common-emitter d.c. output characteristics of an Si n/p/n bipolar junction transistor. (a) The forward configuration and (b) the reverse (backward) configuration.

The output and transfer characteristics in the common-emitter (CE) configuration of an si n/p/n transistor are shown in Figs.732.3 and 732.4 respectively. They are similar to those in the CB configuration given in Figs.732.1 and 732.2 with slight differences in quadrants which should be noted. For example, the saturation region of the forward and reverse CE output characteristics shown in Figs.732.3(a)-(b') are in the first quadrant rather than the second quadrant in the CB configuration because $V_{CE}=V_{BE}-V_{CB}$. Since $V_{CE}>0$ in the saturation range as indicated in Fig.732.3(a), V_{CB} is now negative or the collector-base junction is forward biased which was already noticed in the CB output characteristics shown in Figs.732.1. An important difference is the rather small CE saturation output resistance shown by the large I_C-V_{CE} slope in Fig.732.3(a) while the I_C-V_{CB} slope in Fig.732.1(a) is almost zero (horizontal) in a large portion of the saturation and most of the active regions of the CB configuration.

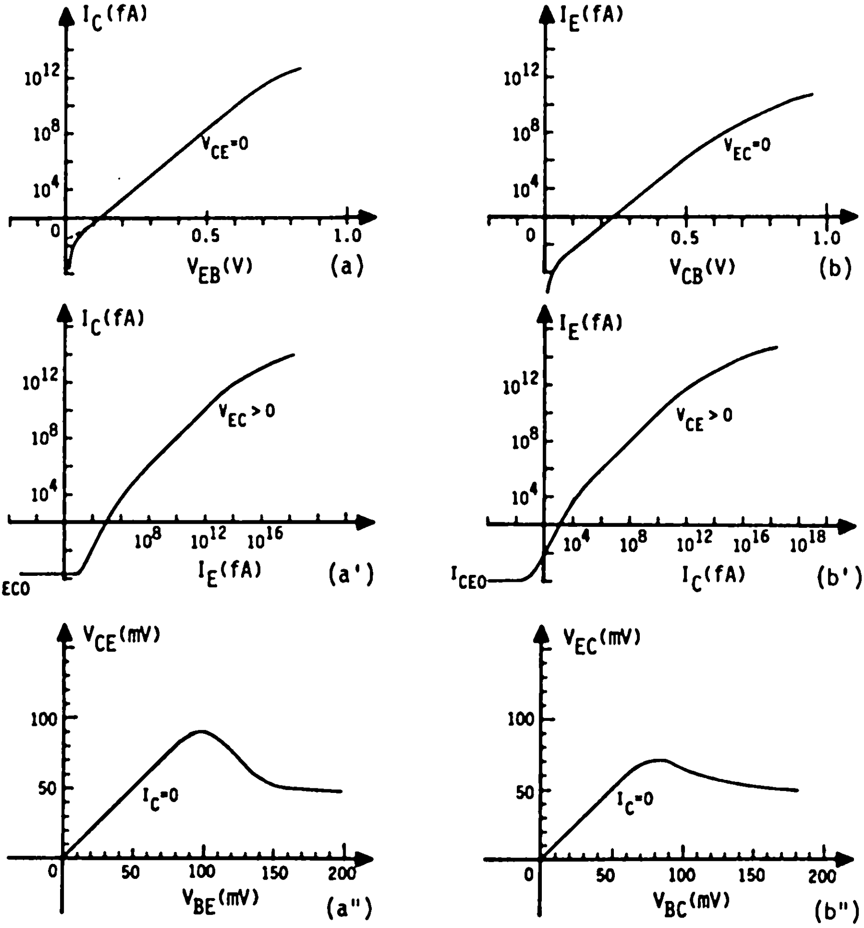


Fig.732.4 The common-emitter d.c. **transfer** characteristics of an Si n/p/n bipolar junction transistor. In the forward configuration, **(a)**, **(a')**, and **(a")**; and the corresponding reverse configuration, **(b)**, **(b')**, and **(b")**.

A similarity between the CE and CB transfer characteristics exists in two of the three characteristics, conductance-transfer and current-transfer shown in Figs.732.4(a)-(a') for forward CE configuration and Figs.732.2(a)-(a') for forward CB configuration, and similarly for corresponding reverse configurations given in the (b)-(b') figures. However, the voltage transfer characteristics of the CE configuration differ significantly from that of the CB configuration as indicated by the (a") and (b") figures. This difference is understandable since the output voltage in the CE configuration, V_{CE}, is the difference of the output and input voltage in the CB configuration, $V_{CE} = V_{CB} - V_{EB}$.

The characteristics of a p/n/p BJT are identical to those just presented for an n/p/n BJT except the sign of the current and voltage are changed. This change of sign rotates all the n/p/n figures by 180 degrees, either clockwise or counter clockwise, to give the corresponding figures of the p/n/p transistor. These electrical characteristics, for n/p/n as well as p/n/p BJTs, will be useful to help explain the basic physics underlying the transistor circuit operation when the algebraic equations of the characteristics are derived and correlated with the material and structure or geometry of the device in the next few sections.

733 Derivation of the D.C. Characteristics of p/n/p BJT

In order to go beyond the simple 2-diode model of the BJT just described, and to give analytical formulae and numerical solutions of the four 2-diode BJT parameters, I_{ED}, I_{CD}, α_F and α_R, we must analyze the transport of electrons and holes in the transistor structure using the six Shockley equations described in chapter 3. The solution would give two equations that relate the two terminal currents (I_E and I_C) to the two terminal voltages (V_{EB} and V_{CB}).

A simple one-dimensional mathematical analysis will be used to obtain the two BJT current-voltage equations. This analysis is necessary to provide a fundamental justification, based on device physics, of the two-port nonlinear network representations of the BJT to be discussed in section 735. The derivation also provides the connection between the two-port circuit parameters and the material and geometrical properties of the transistor structure. This connection is necessary to design transistors to meet a prescribed circuit performance. The simple one-dimensional results can also be used to analyze a real three-dimensional device by dividing (partitioning) the real transistor into regions so that each region can be approximated by the one-dimensional model. Some two- or three-dimensional effects can be approximated using a spatial distribution of several or many one-dimensional transistors.

The figures to be used in the one-dimensional BJT transport analysis are shown in Figs.733.1(a) to (e). These are now explained.

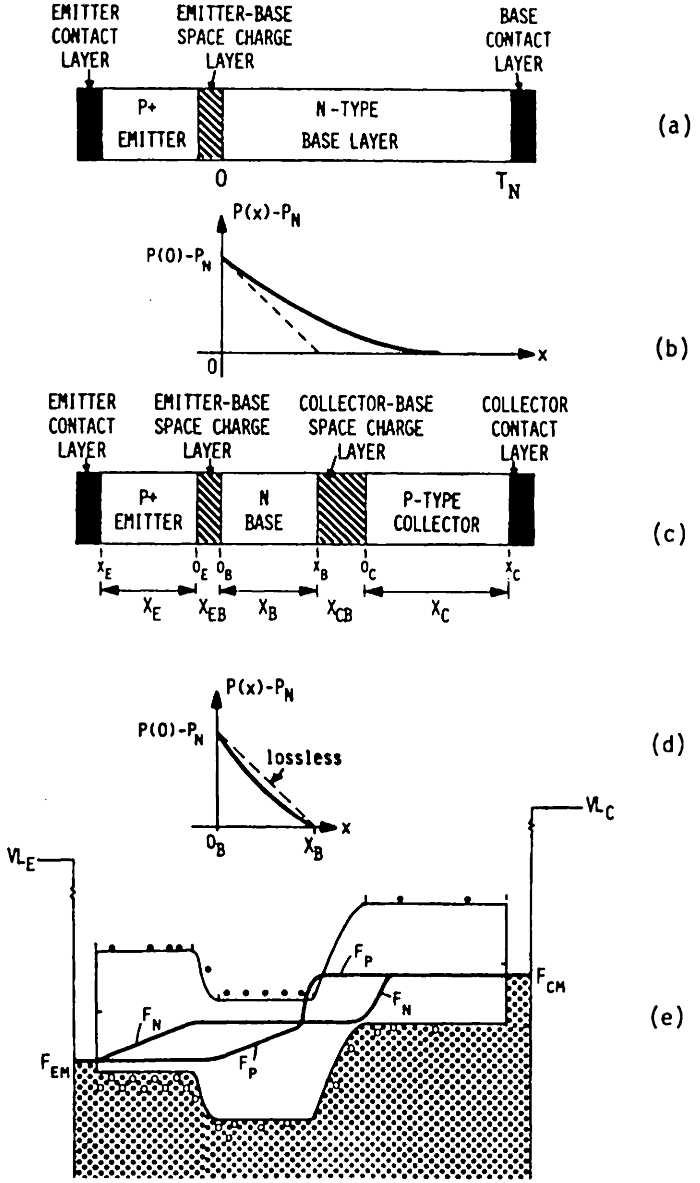

EMITTER CONTACT LAYER
EMITTER-BASE SPACE CHARGE LAYER
BASE CONTACT LAYER
P+ EMITTER
N-TYPE BASE LAYER
0
T_N
(a)
$P(x)-P_N$
$P(0)-P_N$
0
x
(b)
EMITTER CONTACT LAYER
EMITTER-BASE SPACE CHARGE LAYER
COLLECTOR-BASE SPACE CHARGE LAYER
COLLECTOR CONTACT LAYER
P+ EMITTER
N BASE
P-TYPE COLLECTOR
x_E
0_E
0_B
x_B
0_C
x_C
X_E
X_{EB}
X_B
X_{CB}
X_C
(c)
$P(x)-P_N$
$P(0)-P_N$
lossless
0_B
X_B
x
(d)
VL_E
VL_C
F_P
F_N
F_N
F_P
F_{EM}
F_{CM}
(e)

Fig.733.1 Development of the one-dimensional five-layer model of a p/n/p bipolar junction transistor. **(a)** The cross-sectional view of a p+/n diode. **(b)** Minority carrier (hole) distribution in the quasi-neutral base of the p+/n diode. **(c)** The cross-sectional view of the 5-layer p+/n/p BJT with coordinate labels of the three quasi-neutral layers. **(d)** The minority carrier (hole) distribution in the quasi-neutral base layer. **(e)** The energy band diagram showing injected holes (circles in the VB) and electrons (dots in the CB).

The following three assumptions will be made in this one-dimensional BJT model. (1) There are no surface and edge effects. (2) The BJT has two abrupt p/n junctions separated by a thin constant-thickness n-type base layer. And (3), the dopant impurity concentrations are spatially constant in the emitter, base and collector layers separated by the two abrupt p/n junctions.

A slice of a p/n/p BJT is shown in Fig.733.1(c). It is a **five layer boundary value problem**. From the left (emitter) to the right (collector), the five layers are:

(1) The quasi-neutral emitter layer whose parameters are labeled by a single subscript E or e,

(2) The emitter-base junction space-charge layer whose parameters are labeled by a double subscript EB or eb,

(3) The quasi-neutral base layer whose parameters are labeled by a single subscript B or b,

(4) The base-collector or collector-base junction space-charge layer whose parameters are labeled by a double subscript CB or cb.

(5) The quasi-neutral collector layer whose parameters are labeled by a single subscript C or c.

A precise definition and judicial choice of subscripts are crucial in delineating and describing the many physical phenomena that control and limit the current flowing in the five layers. The choice will help to remember the definition, physics, and location of each term in the transistor current equations. The choice is made to avoid duplication of meaning while using a minimum number of subscripts. Thus, the following symbols are unsuitable to designate the parameters and variables of the emitter-base junction space-charge or transition layer. The double subscript et is rejected since t (from transition) is exclusively used for quantities related to a trap or generation-recombination-trapping center in all five layers. The triple and quadruple subscripts esc and escl (from emitter space charge layer) are too long. The double subscript ej (from emitter junction) is ambiguous because the emitter layer is bounded by two junctions on its two surfaces, the metal/p-Si ohmic contact heterojunction and the emitter-base p/n homojunction, so ej could also mean the metal/emitter ohmic contact junction. The **final choice, eb**, has two uniquely useful determinants - it is related to the emitter-base junction or the junction region between the emitter and base quasi-neutral layers, and it indicates a measured quantity or going from the emitter edge to the base edge of the emitter-base junction space-charge layer. Such a judicial thought-out choice is helpful to quickly attain proficiency on learning the complicated device and material physics as well as practicing the even more complex engineering design methodology. It takes many years of experiences to select the optimum symbol and only a very few textbooks written by veteran teacher-researchers (for example D.O.Pederson, J.J.Studer, and J.R.Whinnery, **Introduction to Electronic Systems, Circuits and Devices**, McGraw-Hill, 1966) have even attempted to made a careful selection and have described its pedagogical importance. The lack of good symbols has caused the beginners unwarranted confusion. Poor symbols have made the subject matter and its physics to appear much harder than it actually is. Obscure symbols have made the subject matter and its physics to be much more time-consuming to learn, comprehend, and remember than it should.

We shall now give the derivation of the d.c. current-voltage equations of a p/n/p BJT. We shall first obtain the hole distribution and the hole current as a function of position in the n-type quasi-neutral base layer under the boundary conditions that the emitter-base junction is forward biased, $V_{EB}>0$, and the collector-base junction is short-circuited, $V_{CB}=0$. These junction biases determine the boundary conditions, i.e., the minority carrier concentrations at the boundaries of the two space-charge layers. This approach, i.e., first treating the minority carrier density and current in the base, gives us the fundamental result using a minimum amount of mathematics. Yet the result is so general that it can be and will be used to write down all the minority-carrier current components, in all three quasi-neutral layers and under general bias conditions, without having to solve the differential equation again under the other boundary and junction voltage conditions. For example, it can be used to write down the current flowing in the emitter and collector quasi-neutral layers when the emitter-base junction is short-circuited and the collector-base junction is forward or reverse biased. The applicability of this simple approach has a rigorous mathematical basis: the system is **linear between current density and charge concentration**, and governed by a second-order linear ordinary differential equation of minority carrier diffusion. In a linear system, the currents from two voltage sources or under two boundary conditions are **additive**. This is a tremendous advantage. Although the current-charge relationship is linear, the current-voltage relationship is highly nonlinear since the charge concentration is exponentially dependent on voltage via the Boltzmann factor, exp(qV/kT). The result is a form of the Kirchoff current law in circuit analysis.

The hole distribution in the quasi-neutral base layer can be obtained in the same simple way as that used for the p+/n diode in section 532. The only difference is the boundary condition at the far boundary illustrated in Fig.733.1(a)-(e). For the p+/n diode with a very thick 'base' layer or thick substrate shown in Fig.733.1(a), the equilibrium condition at the far boundary given by

$$P(x=T_N) = P(x=\infty) = P_N$$

was used in section 532 which led to the solutions (532.7n) and (532.8n). This far boundary of the thick n-base layer of the diode diode was allowed to recede to infinity as an approximation in section 532. It was a good approximation even if the physical thickness of the base layer is not very thick as long as it is much thicker than a few minority carrier diffusion lengths in the base, $T_N >> L_p=\sqrt{(D_p\tau_p)}$. For a diffusivity of $D_p=10\text{cm}^2/\text{s}$ and lifetime of $\tau_p=10^{-6}\text{s}$, the diffusion length is $L_p=\sqrt{(10\text{x}10^{-6})}\approx 32\text{x}10^{-4}\text{cm}=32\mu\text{m}$.

For a well-designed high-gain BJT, the opposite is true: the base is very thin. In order to get high amplification or nearly unity alpha, the base layer must be very thin compared with the minority carrier diffusion length so that most of the minority carriers will diffuse through the thin base layer to reach the collector-base junction and few will be lost by recombination with the majority carriers in the base layer. Since equilibrium prevails at the base-side edge of the collector-base junction space-charge layer, X_B in Fig.733.1(c), when $V_{CB}=0$, then the far boundary condition of the thin quasi-neutral base is now

$$P(x=X_B) = P_N. \qquad (733.1)$$

To simplify the algebra and notation, we have used a different coordinate origin for each layer so that the origin coincides with one edge of the layer. To show this, we add a subscript to each of the coordinate origin, 0_E, 0_B and 0_C, as indicated in Fig.733.1(c). The translation of the origin, simplifies the analysis tremendously. Thus, the boundary condition for the concentration of holes injected into the quasi-neutral n-type base layer from its left boundary $x=0_B$ is

$$P(x=0_B) = P_N\exp(qV_{EB}/kT). \qquad (733.2)$$

These two boundary conditions, (733.1) and (733.2), can then be used to solve the minority carrier diffusion equation (532.4) or (532.5)

$$D_p d^2P/dx^2 = (P-P_N)/\tau_p \qquad (733.3)$$

whose general solution, also used in the p/n diode analysis, (532.7), is

$$P(x) - P_N = A\exp(x/L_p) + B\exp(-x/L_p). \qquad (733.4)$$

Using the two boundary conditions given by (733.1) and (733.2), the two constants A and B can be evaluated. After a minor amount of elementary algebra, we get

$$P(x) - P_N = P_N[\exp(qV/kT) - 1]\,\frac{\sinh[(X_B-x)/L_p]}{\sinh[(X_B/L_p)]} \qquad (733.5)$$

which is sketched in Fig.733.1(d) where the lossless case is $L_p>>x_B$.

Thus, the hole current flowing through the baseside boundary of the emitter space-charge layer, $x=0_B$, can be computed from $J_P(x=0)=-qD_pdP(x=0)/dx$. This is the **injected minority carrier** (hole) current from holes injected into the base layer from the emitter layer. Similarly, the injected hole current which reaches the far boundary of the quasi-neutral base layer or the baseside boundary of the collector-base junction space-charge layer, X_B, can be computed from $J_P(x=X_B) = -qD_p dP(x=X_B)/dx$. Using (733.5) for P(x) in the base layer and evaluating the hole current at the two boundaries of the base layer, we have

$$J_P(0_B) = (qD_pP_N/L_p)[\exp(qV_{EB}/kT) - 1]\mathrm{ctnh}(X_B/L_p) \qquad (733.6A)$$

and

$$J_P(X_B) = (qD_pP_N/L_p)[\exp(qV_{EB}/kT) - 1]\mathrm{csch}(X_B/L_p). \qquad (733.6B)$$

The minority carrier diffusion-recombination current in the p-type emitter and collector quasi-neutral layers can also be obtained by solving the linear diffusion equation for the minority carriers (electrons) in these layers. Expressions similar to (733.6A) and (733.6B) can be written down for these two layers without doing any algebra if the boundary conditions at the metal/emitter and metal/collector contacts are the equilibrium condition given by (733.1). This equilibrium boundary condition assumes that the contact between the collector metal terminal and the collector semiconductor body or silicon substrate is perfect (i.e. zero resistance and zero voltage drop at all current levels). Similarly, the emitter contact is assumed to have a zero contact resistance. These zero resistance contacts imply that the electron and hole lifetimes are zero, or the electron and hole recombination and generation rates at these contacts are infinite at all current densities. Finite contact resistance exists in real situations. Contact resistance calculations and examples of reducing or minimizing contact resistances were described in sections 58n.

To put this equilibrium boundary condition at the contacts on a physical and mathematically more rigorous basis, consider the metal/p-Si contacts to the emitter and collector of the p/n/p transistor. The 1-d (one-dimensional) d.c. continuity equation for the minority carriers (electrons in these p-type regions) can be obtained in a similar way as that obtained for holes in the n-type base which was given by (532.3) in the one-junction diode analysis. For the electrons in the p-type emitter and collector layers, we have

$$dJ_N/dx = -q(G_N-R_N) = -q(N_P-N)/\tau_n. \qquad (733.7)$$

The perfect (zero-resistance) contact between an n-type semiconductor and the metal layer is defined by $\tau_n=0$ at this contact interface. Thus, (733.7) shows that $N=N_P$, i.e., electron concentration at the contact must be equal to its thermal equilibrium value. This is necessary so that dJ_N/dx at the contact may have a non-zero and non-infinity value at the contact boundary.

In the case of a metal/n-semiconductor contact, the current on both sides of the contact are carried by electrons. Thus, the electron current density is continuous through the contact or $dJ_N/dx=0$, and J_N(to the left of the contact) = J_N(to the right of the contact).

For a metal/p-semiconductor contact of the p/n/p transistor, there is a discontinuity of the hole current at the metal/p-Si-emitter and p-Si-collector/metal contacts since the current in the metal is carried by electrons while it is carried by holes (majority carriers) in the p-type semiconductor. The continuity of the total current is maintained which can be demonstrated using the Kirchoff current law which states that the total current must be constant in a one-dimensional problem, i.e., $J=J_N(x)+J_P(x)=\text{constant}$ through a contact or any cross-sectional area in all one-dimensional problems.

If the contact is not perfect, i.e., $\tau_n\neq0$ and $\tau_p\neq0$ at the contact, then the boundary condition at the contact is more complicated. In such a nonideal contact, instead of using the lifetime to specify the 'goodness' of a contact, a **surface recombination velocity** or more precisely, **interfacial recombination velocity**, has been used. The concept of surface recombination velocity can be mathematically derived and defined using the continuity equations such as (733.7) and (532.3). It can also be derived and defined by considering the boundary as a thin interfacial layer with a finite thickness, and then letting the thickness approach zero to define an average recombination-generation rate which is a sum of the distribution of the recombination-generation rates through the thickness of the interfacial layer. This was described in section 582.

In our simple model as the first encounter for the beginner, we will assume that the metal/p-Si contacts to the p-type emitter and p-type collector surfaces are ideal (zero resistance) contacts. Then we have thermal equilibrium or $N=N_P$ at these contacts. This is the same boundary condition we had for holes at the collector-base junction $[P(x=X_B)=P_N]$ given by (733.1). Thus, the diffusion equation for the minority carriers, electrons in the p-type emitter and collector quasi-neutral layers, given by

$$D_n d^2N/dx^2 = (N - N_P)/\tau_n \tag{733.8}$$

would have exactly the same solution as that of the holes in the n-type base which were given by (733.6A) and (733.6B). To use these earlier hole solutions for electrons, all we have to do is to interchange N and P and n and p. But, we have two p-type layers: the p-type emitter and the p-type collector layers. Thus, we will have some notation confusion if we use the subscripts N and P to denote the quantities in the two layers such as N_P as the thermal equilibrium minority carrier concentration (electron) in the two p-type layers. To avoid the confusion, we note that we are dealing with **minority carrier** diffusion-recombination currents in these three layers. Thus, if we use the subscript E, B and C for the **minority carrier** quantities in the three layers, we can simplify and unify the notation. For example, the minority carrier (hole) currents at the two boundary surfaces of the base layer, (733.6A) and (733.6B), can be rewritten as

$$J_P(0_B,V_{EB}) = (qD_BP_B/L_B)[\exp(qV_{EB}/kT) - 1]\mathrm{ctnh}(X_B/L_B) \qquad (733.9A)$$

and

$$J_P(X_B,V_{EB}) = (qD_BP_B/L_B)[\exp(qV_{EB}/kT) - 1]\mathrm{csch}(X_B/L_B) \qquad (733.9B)$$

where D_B = minority carrier diffusivity in the base layer = D_p, $P_B=P_N$ and $L_B=L_p=\sqrt{D_B\tau_B}=\sqrt{D_p\tau_p}$. Note, the key of notation simplification and unification is that they all refer to the **minority carriers**, i.e. holes in the n-type base layer.

With this notation convention, we can then write down the electron (minority carrier) currents flowing through the two surfaces of the quasi-neutral emitter layer which is assumed to have a thickness of X_E:

$$J_N(0_E,V_{EB}) = (qD_EN_E/L_E)[\exp(qV_{EB}/kT) - 1]\mathrm{ctnh}(X_E/L_E) \qquad (733.10A)$$

$$J_N(X_E,V_{EB}) = (qD_EN_E/L_E)[\exp(qV_{EB}/kT) - 1]\mathrm{csch}(X_E/L_E) \qquad (733.10B)$$

There is no minority carrier current flowing in the collector layer since so far we have short circuited the collector-base junction or collector-base terminals (assuming no contact and bulk resistances so the junction and terminal voltages are equal and equal to zero in this BJT analysis methodology.) If a d.c. voltage, V_{CB}, is applied across the collector-base junction, then we will have four current components similar to those given by (733.9A) to (733.10B) due to the applied voltage, V_{CB}. The solutions can be immediately written down like (733.9A) to (733.10B) without solving the differential equations since the boundary conditions are the same. The currents due to V_{CB} are then

$$J_P(0_B,V_{CB}) = -\ (qD_BN_B/L_B)[\exp(qV_{CB}/kT) - 1]\mathrm{csch}(X_B/L_B) \qquad (733.11A)$$

$$J_P(X_B,V_{CB}) = -\ (qD_BN_B/L_B)[\exp(qV_{CB}/kT) - 1]\mathrm{ctnh}(X_B/L_B) \qquad (733.11B)$$

$$J_N(X_B,V_{CB}) = -\ (qD_CN_C/L_C)[\exp(qV_{CB}/kT) - 1]\mathrm{ctnh}(X_C/L_C) \qquad (733.12A)$$

$$J_N(0_C,V_{CB}) = -\ (qD_CN_C/L_C)[\exp(qV_{CB}/kT) - 1]\mathrm{csch}(X_C/L_C) \qquad (733.12B)$$

The Shockley BJT Equations

The total current per unit area flowing into the emitter terminal or through the emitter quasi-neutral and emitter-base junction space-charge layers is just the sum of the three terms at the two edges of the emitter-base junction space-charge layer given respectively by (733.10A), (733.9A), and (733.11A):

$$J_E = +\ J_N(0_E,V_{EB}) + J_P(0_B,V_{EB}) + J_P(0_B,V_{CB}) \qquad (733.13)$$

$$\begin{aligned} &= +\ (qD_EN_E/L_E)\mathrm{ctnh}(X_E/L_E)[\exp(qV_{EB}/kT)-1] \\ &\quad +\ (qD_BP_B/L_B)\mathrm{ctnh}(X_B/L_B)[\exp(qV_{EB}/kT)-1] \\ &\quad -\ (qD_BN_B/L_B)\mathrm{csch}(X_B/L_B)[\exp(qV_{CB}/kT)-1] \end{aligned} \qquad (733.14)$$

Similarly, the total current per unit area flowing into the collector terminal or through the collector quasi-neutral and collector-base junction space-charge layers is the sum of the three currents at the two edges of the collector-base junction space charge layer, (733.9B), (733.11B) and (733.12A). The collector current flowing into the p-type collector body or collector quasi-neutral layer from the collector metal terminal or the external circuit is directed in the **negative x-axis direction** in the coordinate system we used to solve the diffusion equation. Thus, we add a negative sign and get

$$J_C = -J_P(X_B,V_{EB}) - J_P(X_B,X_{CB}) - J_N(0_C,V_{CB}) \qquad (733.15)$$

$$\begin{aligned} &= -(qD_BN_B/L_B)\mathrm{csch}(X_B/L_B)[\exp(qV_{EB}/kT)-1] \\ &\quad + (qD_BN_B/L_B)\mathrm{ctnh}(X_B/L_B)[\exp(qV_{CB}/kT)-1] \\ &\quad + (qD_CN_C/L_C)\mathrm{ctnh}(X_C/L_C)[\exp(qV_{CB}/kT)-1] \end{aligned} \qquad (733.16)$$

The two current equations given by (733.14) and (733.16) were first derived (in different notation and slightly simplified form) by Shockley in 1949 [733.1] when he theoretically invented the bipolar junction transistor and published the detailed analyses. The pair has been known as the **Shockley Bipolar Junction Transistor Equations**. These two equations can be readily identified with the two current equations of the 2-Diode BJT model given by (731.1) and (731.2).

[733.1] W.Shockley, "Theory of p-n junctions in semiconductors and p-n junction transistors," Bell Syst. Tech. J. 28(7), pp.435-489, July 1949.

[733.2] C.T.Sah, R. N. Noyce and W. Shockley, "Carrier generation and recombination in p-n junction and p-n junction characteristics," Proc.IRE 45(9), pp.1228-1248. Sept. 1957.

[733.3] C.T.Sah, "A new semiconductor tetrode, the surface-potential controlled transistor," Proc.IRE, 49(11), pp.1623-1634, November 1961.

[733.4] C.T.Sah, "Effect of surface recombination and channel on p-n junction and transistor characteristics," IRE Trans. Electron Devices, ED-9(1), pp.94-108, January, 1962.

The SNS BJT Equations

The Shockley BJT equations contain only the diffusion-recombination currents from the three quasi-neutral layers of the five-layer BJT. The recombination-generation currents from the two space-charge layers were not included by Shockley in his 1949 transistor equations although in that paper he had given a simplified description of the effects from a thin-layer of very-high-density recombination centers in the space-charge layer on the diode characteristics. When Si diodes and transistors became commercially available in the mid-1950's, these recombination-generation currents from the space-charge layer were found to be important and even dominant in the low forward-current and the entire reverse-current ranges of Si BJTs. In sections 535 and 537 on p/n junction diodes, we have demonstrated the importance of recombination-generation currents in the space-charge layer of Si p/n diodes. The 1949 Shockley BJT equation was extended in 1957 [733.2] by Sah, Noyce and Shockley to include the recombination-generation current in the space-charge layer. They derived a general diode current-voltage formulae whose simplified form was derived in (535.4). Extensive correlations of the SNS diode formulae with experimental data were given by Sah in 1961 [733.3] and 1962 [733.4]. Selected theoretical-experimental comparison examples from these two articles were presented in section 537.

The Shockley BJT equations can be similarly extended to include the recombination-generation currents in the emitter-base and collector-base junction space-charge layers. To implement this extension, we use the SNS diode formulae given by (535.4) and change the symbols to adapt them to the emitter-base (subscript EB) and collector-base (subscript CB) junction space-charge layers. These space-charge layer currents are

$$J_{EB}(V_{EB}) = (qn_i/\tau_{EB})X_{EB}[\exp(qV_{EB}/2kT) - 1] \qquad (733.17)$$

and

$$J_{CB}(V_{CB}) = (qn_i/\tau_{CB})X_{CB}[\exp(qV_{CB}/2kT) - 1] \qquad (733.18)$$

Then, J_{EB} can be added to J_E of (733.14), and J_{CB} to J_C of (733.16), to give the extended Shockley BJT equations or the SNS BJT equations which are

$$\begin{aligned} J_E &= J_E(\text{Shockley}) + J_{EB}(V_{EB}) \\ &= + J_N(0_E,V_{EB}) + J_{EB}(V_{EB}) + J_P(0_B,V_{EB}) + J_P(0_B,V_{CB}) \\ &= + (qD_EN_E/L_E)\text{ctnh}(X_E/L_E)[\exp(qV_{EB}/kT) - 1] \\ &\quad + (qn_iX_{EB}/\tau_{EB})[\exp(qV_{EB}/2kT) - 1] \\ &\quad + (qD_BP_B/L_B)\text{ctnh}(X_B/L_B)[\exp(qV_{EB}/kT) - 1] \\ &\quad - (qD_BN_B/L_B)\text{csch}(X_B/L_B)[\exp(qV_{CB}/kT) - 1] \end{aligned} \qquad (733.19')$$

and

$$\begin{aligned} J_C &= J_C(\text{Shockley}) + J_{CB}(V_{CB}) \\ &= - J_P(X_B,V_{EB}) - J_P(X_B,V_{CB}) + J_{CB}(V_{CB}) - J_N(0_C,V_{CB}) \\ &= - (qD_BN_B/L_B)\text{csch}(X_B/L_B)[\exp(qV_{EB}/kT) - 1] \\ &\quad + (qD_BN_B/L_B)\text{ctnh}(X_B/L_B)[\exp(qV_{CB}/kT) - 1] \\ &\quad + (qn_iX_{CB}/\tau_{CB})[\exp(qV_{CB}/2kT) - 1] \\ &\quad + (qD_CN_C/L_C)\text{ctnh}(X_C/L_C)[\exp(qV_{CB}/kT) - 1] \end{aligned} \qquad (733.20')$$

These can be put into the following compact form which is a facsimile of the two current equations in the 2-Diode BJT model given by (731.1) and (731.2):

$$J_E = +\,(j_E+j_{EB}+j_B)[\exp(qV_{EB}/kT)-1] - (\alpha_B j_B)[\exp(qV_{CB}/kT)-1] \qquad (733.19)$$

$$J_C = -\,(\alpha_B j_B)[\exp(qV_{EB}/kT)-1] + (j_C+j_{CB}+j_B)[\exp(qV_{CB}/kT)-1] \qquad (733.20)$$

The SNS current coefficients in the above equations are defined by

$$j_E = (qD_E N_E/L_E)\mathrm{ctnh}(X_E/L_E) \qquad (733.21A)$$

$$j_{EB} = (qn_i X_{EB}/\tau_{EB})[\exp(qV_{EB}/2kT)-1]/[\exp(qV_{EB}/kT)-1]$$

$$= (qn_i X_{EB}/\tau_{EB})[\exp(qV_{EB}/2kT)+1] \qquad (733.21B)$$

$$j_B = (qD_B P_B/L_B)\mathrm{ctnh}(X_B/L_B) \qquad (733.21C)$$

$$\alpha_B = \mathrm{sech}(X_B/L_B) \qquad (733.21D)$$

$$j_C = (qD_C N_C/L_C)\mathrm{ctnh}(X_C/L_C) \qquad (733.21E)$$

and

$$j_{CB} = (qn_i X_{CB}/\tau_{CB})[\exp(qV_{CB}/2kT)-1]/[\exp(qV_{CB}/kT)-1]$$

$$= (qn_i X_{CB}/\tau_{CB})[\exp(qV_{CB}/2kT)+1] \qquad (733.21F)$$

Notice that we are using the symbol α_B in (733.21D) to emphasize that it is due to base recombination. Original (1950's and 1960's) and recent literature has used α_T (T=base transport) and β^* which obscure and confuse the physics.

Numerical illustrations on the importance of the recombination-generation currents in the junction space-charge layer, discussed in sections 535 and 537, are applicable to BJT. Additional numerical examples are given in section 738.

In the preceding derivation, no reference was made to the energy band diagram shown in Fig.733.1(e) which can be helpful to illustrate the flow of the holes from the p+ emitter through the base to the p-type collector. The energy band diagram was used in chapter 5 to derive the three components of the p/n junction diode current. A shortcut has been taken here for the BJT by using the diode results to bypass all of the tedious and obscuring algebra traditionally used to derive the Shockley transistor equations and the rarely presented derivations of the SNS transistor equations. This shortcut not only focuses the description on the basic device physics and theory of the BJT, but also provides a simple and clear mental image of the algebra without having to write down each step explicitly.

734 The Original and Extended Ebers-Moll Equations of BJT

The total emitter-base and collector-base junction currents in a real transistor can be obtained only approximately from multiplying the current density equations of the one-dimensional Shockley-SNS transistor, (733.19) and (733.20) by the emitter-base and collector-base junction areas respectively because of two- and three-dimensional effects from the unequal emitter and collector junction areas and the series emitter, base and collector resistances which cause current crowding near the edge of the emitter-base and collector-base junctions. The emitter-base junction area, A_E, is usually smaller than the collector-base junction area, A_C, in order to allow for ohmic contact to be made to the base layer. The cross-sectional view of the double-diffused Si planar BJT shown in Figs.720.1(b), 720.2(b) and 720.3(n) illustrates $A_C > A_E$. Thus, the collector current equation, (733.20), must be extended to include the current flowing through the extended portion of the collector-base junction area, A_C-A_E, which is not overlapped by the emitter-base junction. This additional collector/base diode is known as the **not-overlapped collector-base diode** (or **underlap collector-base diode**). To include the effects of this underlap collector-base junction diode, a p/n diode is added between the collector and base terminals whose area is A_C-A_E. There are two differences between this underlap collector-base junction diode and the emitter-overlapped collector-base junction diode of a p/n/p transistor. (1) The n-type base layer of this underlap diode, $X_{B\text{-}ul} = X_E + X_{EB} + X_B$, is thicker than the n-type base of the p/n/p transistor, X_B. In some BJT designs, $X_{B\text{-}ul}$ of the underlap diode may be thinner if an n+ surface ring [like the circular n/p/n shown in Fig.720.1(b) and other non-circular geometry] is diffused into the exposed n-base surface to provide low resistance ohmic contact between the base metal and the n-type base. The n+ base surface ring is narrower than the ring region whose area is A_C-A_E. (2) Even without the n+ contact ring, part of the top surface of this underlap diode is oxide passivated with very low interface recombination rate instead of the infinite recombination rate at the base metal contact.

Thus, the collector current equation of a real transistor, such as the planar transistor just described, consists of two parts: a transistor collector current of the **intrinsic transistor** with an effective collector area approximately equal to the emitter area, A_E, and a diode current from the **underlap** or **non-overlap collector-base diode** with an effective area, A_C-A_E. In the active forward mode, the underlap collector-base diode current is just the reverse leakage current of the n/p junction which increases the collector current by slightly less than the area ratio $(A_C$-$A_E)/A_E$. It is less since part of the underlap area is oxide passivated instead of completely covered by a base contact metal. However, the underlap collector diode becomes important when the BJT is in the saturation mode and is to be switched off from the saturation mode since both the emitter and the collector-base junctions are forward biased and injecting minority carriers into the base layer. These stored excess minority carriers must be eliminated from the base in order to return the collector-base junction to the low-current reverse-biased state when the transistor is switched off from the saturation mode. But the underlap collector diode does not have a short-circuited emitter-base junction to speed up the extraction of the stored excess minority (holes) carriers previously injected during the forward bias on the collector-base junction in the saturation mode. This will make the delay in the collector current waveform higher than that of an intrinsic transistor.

Reduction or elimination of the non-overlap or underlap collector-base junction area was the key objective in developing the latest state-of-the-art submicron picosecond silicon BJT transistors for the CPU chips of mainframe and supercomputers in the mid-1980's. The IBM team (Hwa N. Yu, Tak H. Ning, Denny Tang, ...) successfully developed such a technology, known as the self-aligned emitter first envisioned by Japanese engineers in the mid-1970's, using: heavily As-doped poly-Si emitter as the contact, the As in the poly-Si emitter contact as the n-type impurity diffusion source, and the self-aligned mask for a p+ base-contact ring. The reduction of the non-overlap collector diode also reduces the lateral base resistance significantly. Designing the BJT geometry to give a minimum underlap collector diode area is the crucial step to increase the switching speed to approach the intrinsic switching speed of an ideal BJT. A different approach was used by Texas Instruments in the mid-1950's to manufacture a line of logic gates using Si BJTs. In this approach, a Schottky (Bethe) diode was connected across the collector-base junction. It serves two purposes: (i) to sink all the collector saturation current since the Bethe-Schottky Al/p-Si or Al/n-Si heterojunction diode will conduct heavily at a lower forward bias than the collector-base p/n Si homojunction because the Schottky-Bethe diode has a lower **cut-in voltage** (see section 563), and (ii) to limit the forward bias voltage across the collector-base junction to a low forward value thereby preventing the CB homo Si junction from heavily injecting minority carriers into the quasi-neutral base and collector layers. A schematic of the Schottky barrier clamped BJT was shown in Fig.565.2. This was a temporary solution popularly used by Texas Instruments, Motorola, and other manufacturers in the 1970's to produce small-scale integrated (SSI <1000 transistors per chip) bipolar circuits, such as the Schottky-clamped Transistor-Transistor Logic (TTL) gate arrays. It gave Si logic gates with lower power dissipation and higher speed. The Schottky-clamped BJTs did not reach the intrinsic speed of the transistor because the Schottky diode (i) increases the collector output capacitance from the non-active collector-base junction area, and (ii) does not decrease the lateral series resistance of the thin base layer.

In order to understand the characteristics of a real BJT, the 1-d **ideal** or **intrinsic transistor** is first studied in this section. The 2-d parasitics that cannot be ignored in real transistors, such as the non-overlap diode, the lateral base resistance, and the series bulk and contact resistances, are studied in a following section, 736.

The ideal intrinsic BJT has equal emitter-base and collector-base junction areas. For the double diffused BJT shown in Fig.720.3(n), this is the portion under the emitter-base junction. The total emitter and collector currents of the intrinsic BJT, I_E and I_C, are obtained from multiplying the current densities, J_E and J_C given by (733.19) and (733.20), by the emitter area, A_E, since the collector-base junction of the intrinsic transistor is that portion of the collector area which is covered by the emitter-base junction. The total current equations are then given by the following SNS transistor equations:

$$I_E = J_E A_E$$
$$= + (j_E + j_{EB} + j_B) A_E[\exp(qV_{EB}/kT)-1] - \alpha_B j_B A_E[\exp(qV_{CB}/kT)-1] \qquad (734.1)$$

$$I_C = J_C A_E$$
$$= - \alpha_B j_B A_E[\exp(qV_{EB}/kT)-1] + (j_C + j_{CB} + j_B) A_E[\exp(qV_{CB}/kT)-1] \qquad (734.2)$$

They can be written in a compact form, introduced by Ebers and Moll in December, 1954, and known as the **Ebers and Moll BJT equations**. In this book, we call them the **extended Ebers and Moll (xE-B) BJT equations** since the original Ebers and Moll equations neglected the recombination-generation currents in the two space-charge layers, i.e. by letting $j_{EB}=j_{CB}=0$ in the above equations. The xE-B equations are:

$$I_E = + \ I_{ES}[\exp(qV_{EB}/kT) - 1] - \alpha_R I_{CS}[\exp(qV_{CB}/kT) - 1] \qquad (734.3)$$

$$I_C = - \alpha_F I_{ES}[\exp(qV_{EB}/kT) - 1] + \ I_{CS}[\exp(qV_{CB}/kT) - 1] \qquad (734.4)$$

The SNS and EM parameters are related and defined by the following list.

$$I_{ES} = A_E(j_E + j_{EB} + j_B) \qquad (734.5)$$

$$I_{CS} = A_E(j_C + j_{CB} + j_B) \qquad (734.6)$$

$$\alpha_R = A_E(\alpha_B j_B/I_{CS}) = \alpha_B[j_B/(j_C+j_{CB}+j_B)] \equiv \alpha_B \gamma_C \qquad (734.7A)$$

$$\alpha_F = A_E(\alpha_B j_B/I_{ES}) = \alpha_B[j_B/(j_E+j_{EB}+j_B)] \equiv \alpha_B \gamma_E \qquad (734.7B)$$

$$\alpha_B = \operatorname{sech}(X_B/L_B) \qquad (734.8)$$

$$\gamma_C = j_B/(j_C+j_{CB}+j_B) \qquad (734.9A)$$

$$\gamma_E = j_B/(j_E+j_{EB}+j_B) \qquad (734.9B)$$

The factor, α_B, is known as the **base transport factor**. It is the fraction of the injected minority carriers that have escaped recombination during the diffusion-drift transport through the base layer and are collected by the other junction. For a real transistor such as the diffused base planar transistor, built-in electric field is present in the base layer due to the base impurity concentration gradient. This field would aid the minority carrier diffusion from the emitter to the collector (the forward direction) but retard the diffusion in the reverse direction. Thus, a forward base transport factor, α_{BF}, would appear in the forward alpha, α_F, and a different reverse base transport factor, α_{BR}, would appear in the reverse or inverse alpha, α_R. These have the same form as (734.8) but with an additional field acceleration or retardation factor in the argument, η_F or η_R, due to the electric field in the base layer. The field acceleration or retardation factor appears as follows.

$$\alpha_{BF} = \text{sech}(\eta_F X_B/L_B) \tag{734.10A}$$

$$\alpha_{BR} = \text{sech}(\eta_R X_B/L_B) \tag{734.10B}$$

A similar directional dependence from unsymmetrical or dissimilar emitter, base and collector dopant impurity concentration profiles would also appear in j_B which is the coefficient of the injected minority carrier current into the base.

The gamma's, γ_E and γ_C, are known as the d.c. minority carrier **injection efficiency** of the emitter-base and collector-base junctions. Because they are specific to the junctions, the subscripts E and C are not replaced by F and R which would have caused confusion. The physics of γ is easy to understand and remember. For example, $\gamma_E = j_B/(j_E + j_{EB} + j_E)$, is the fraction of the total emitter-base junction current due to minority carriers injected into the base (the j_B part). The other two components, j_{EB} and j_E, are respectively the recombination current in the emitter-base junction space-charge-layer and the minority-carrier injection current into the emitter layer and hence they are lost or not available for collection by the collector-base junction. Thus, j_{EB} and j_E reduce the minority-carrier injection efficiency of the emitter-base junction. A similar interpretation can be given to the injection efficiency of the collector-base junction, γ_C.

The Ebers-Moll equation can be compared with the 2-diode transistor equations, (731.1) and (731.2) repeated below

$$I_C = I_{CD} - \alpha_F I_{ED}. \tag{731.1}$$

$$I_E = I_{ED} - \alpha_R I_{CD} \tag{731.2}$$

This shows that the parameters in the 2-diode equations are related to the EM parameters as follows.

$$I_{ED} = I_{ES}[\exp(qV_{EB}/kT) - 1] \tag{734.11A}$$

and

$$I_{CD} = I_{CS}[\exp(qV_{CB}/kT) - 1]. \tag{734.11B}$$

As indicated earlier, (734.1) and (734.2) reduce to the **original Ebers-Moll Equations** and the SNS parameters given in (734.5) to (734.10B) reduce to the EB parameters when the recombination-generation currents in the emitter-base and collector-base junction space-charge layers are deleted by setting $j_{EB}=0$ and $j_{CB}=0$ in (734.5) to (734.10B). This deletion is a good approximation when the current is high or when the forward junction bias is above the injection threshold voltage defined by (535.7), about $V_{BE} > 0.3\text{V}$, so that recombination in the junction space-charge layer can be neglected. However, the generation current in the space-charge layer of the reverse-biased collector-base junction cannot be neglected since it dominates at all collector-base junction voltages. But this would modify the original Ebers-Moll equation only slightly, giving the BJT a finite output conductance rather than zero as predicted by the saturation of the reverse current in the Shockley diode equation. Including these space-charge-layer currents as we have done in the definitions given in (734.5) to (734.10B), we then have the **Extended Ebers-Moll equations, x-EM equations.**

When the extended Ebers-Moll d.c. transistor equations given by (734.1) and (734.2) are plotted, they give accurate prediction of the silicon transistor characteristics in all four operation regions (the forward and reverse active, the cutoff, and the saturation regions) at low voltages and low currents. The low voltage limitation comes from the fact that we have not included interband impact generation of electron-hole pairs in the collector-base junction space-charge layer which would occur at large reverse collector-base bias voltages. The low current limitation comes from the fact that we have assumed: a linear recombination law, small injected excess minority carrier density compared with majority carrier density, dominant diffusion current, and negligible drift current. The two-dimensional effects are unimportant at low currents and voltages and they are also unimportant in well-designed transistors. The residual two- and three-dimensional effects at low current levels in a real transistor can be approximated via partitioning by **dissecting** the transistor into several one-dimensional transistors and diodes and by applying the one-dimensional diode and transistor models just presented to each of these **regional diodes and transistors**. The high current level effects and their onset conditions are described and analyzed in section 738. The high voltage effects are analyzed in section 739.

The principal improvement of the extended Ebers-Moll model over the original Ebers-Moll equations occurs at the low current or low emitter-base junction voltages range when electron-hole recombination in the emitter-base junction space-charge layer is important. This recombination loss reduces the amplification and gain of the BJT at low current levels. It has strongly influenced the low-power and high-impedance analog circuit design in the history of bipolar integrated circuit advancement. For example, the high input-impedance integrated bipolar operational amplifier circuit chip was successfully designed by Widler and volume produced by the National Semiconductor Corporation in the mid-1960s owing to their ability to fabricate a BJT having high gain and high input impedance at very low base and collector currents, known as the super-beta transistor, with reported current gains exceeding 10,000 at low currents. We shall postpone to a later section the analysis of the material and geometrical parameters required to design high performance BJT since many interesting transistor properties in circuit applications can be readily obtained from the Ebers-Moll model without additional knowledge of the Ebers-Moll and extended Ebers-Moll or SNS parameters. These analyses are traditionally given in the beginning sections of introductory digital circuit textbooks and courses. Their analyses and results are directly applicable to the extended Ebers-Moll model which we just described based on the SNS transistor equations.

735 Two-Port Nonlinear D.C. Network Representations of BJT

The Ebers-Moll d.c. current-voltage equations given by (734.1) and (734.2) have been used to analyze and design BJT circuits. It is one of the six pairs of **two-port d.c. network equations** relating the four variables: the input and output currents and charges (instead of voltages). The charges are proportional to the voltages via the Boltzmann factor, exp(qV/kT), which is the fundamental cause of nonlinearity in a BJT. The linear d.c. two-port network representation of a three-terminal or four-terminal network gives two equations relating the four terminal variables: the two terminal currents, I_1 $(=-I_3)$ and I_2 $(=-I_4)$, and the two terminal voltages, V_{13} and V_{24}. The traditional three-terminal (or four-terminal if terminals 3 and 4 are not common) two-port d.c. network representation for linear circuits must be generalized to represent the **nonlinear** two-port d.c. networks of the BJT if d.c. voltages instead of d.c. charges are used to give the four variables, V_{EB}, V_{CB}, I_E and I_C, in the common-base configuration.

The EM equations were used to relate the **two-port d.c. circuit parameters** to the material and geometry parameters of the intrinsic Shockley transistor. These relationships have been given in introductory circuit textbooks and handbooks as the basis for the design methodologies of digital and analog bipolar integrated circuits. Without reworking the algebra, the methodologies and results can be directly applied to the SNS transistor equations or the extended Ebers-Moll model which include recombination-generation in the two space-charge layers since the SNS or extended EB equations have the same form as the original EM equations. Only the material parameters need to be extended in order to apply these 2-port results to a Si transistor at low current using the SNS or x-EB model.

In this section, we generalize the analysis of the three-terminal BJT by deriving additional three-terminal two-port network equations of the BJT from the EM equations. We then deduce the simplified two-port equations in the four regions of the current-voltage characteristics (or modes of operation) in which the BJT is biased into: the **forward active mode**, the **reverse active mode**, the **cutoff mode** and the **saturation mode**. We shall consider both the common-base and common-emitter configurations. We will also give the simplified two-port equivalent-circuit models of the BJT in each of these four modes of operation and in the two (CB and CE) circuit configurations.

The p/n/p transistor will be analyzed. The results are directly applicable to n/p/n transistors with appropriate change of sign of the voltages and current coefficients. In the next two subsections, we shall first derive the alternative general two-port equations from the Ebers-Moll equations in the common-base and common-emitter configurations. These alternative general equations will expedite the algebra required to derive the simplified equations in the four modes of BJT operation. The algebra is simple and straightforward, involving only the inversion of the original common-base Ebers-Moll equation written as a 2x2 matrix equation.

General Common-Base Two-Port Network Equations

The common-base equations in the preceding section, (734.1) - (734.10B), will be used as the starting point to obtain the other two-port network equations and parameters. The common-base EM equations given by (734.1) and (734.2) have the terminal voltages, V_{EB} and V_{CB}, as the independent variables and the two terminal currents, I_E and I_C, as the dependent variables. This pair comprises the **nonlinear short-circuit-parameter** or **nonlinear y-parameter two-port network equations** given below. However, in some of the following analyses it is more convenient to use alternative pairs among the four independent variables.

$$I_E = I_{ES}[\exp(qV_{EB}/kT) - 1] - \alpha_R I_{CS}[\exp(qV_{CB}/kT) - 1] \quad (735.1)$$

and

$$I_C = -\alpha_F I_{ES}[\exp(qV_{EB}/kT) - 1] + I_{CS}[\exp(qV_{CB}/kT) - 1] \quad (735.2)$$

where I_{ES}, I_{CS}, α_F and α_R are the four **d.c. short-circuit two-port parameters.** They are obviously not the admittances or conductances of the linear short-circuit or y-parameter two-port network due to the explicit nonlinearity from the Boltzmann factors, $\exp(qV_{EB}/kT)$ and $\exp(qV_{CB}/kT)$, and also the implicit nonlinearity in I_{ES} and I_{CS} due to recombination-generation currents in the two junction space-charge layers and other voltage and current modulations of the geometrical and material parameters of the BJT which are discussed in the following sections.

The short-circuit two-port equations given above can be more simply written in terms of the emitter and collector diode current equations following the definitions of the two-diode model given by (731.1) and (731.2)

$$I_E = I_{ED} - \alpha_R I_{CD} \quad (735.1A)$$

and

$$I_C = I_{CD} - \alpha_F I_{ED}. \quad (735.2A)$$

The diode currents above have the simple form of the Shockley diode equation

$$I_{ED} = I_{ES}[\exp(qV_{EB}/kT) - 1] \quad (735.3)$$

and

$$I_{CD} = I_{CS}[\exp(qV_{CB}/kT) - 1]. \quad (735.4)$$

However, the coefficients, I_{ES} and I_{CS}, are extended, as indicated in (734.5) and (734.6), to include the recombination-generation currents in the space-charge layers initially derived for the SNS diode equation in section 535.

Inverting the above pair to make the terminal currents, I_E and I_C, as the independent variables, we get the **open-circuit two-port equations:**

$$(1 - \alpha_F\alpha_R)I_{ES}[\exp(qV_{EB}/kT) - 1] = I_E + \alpha_R I_C \tag{735.5}$$

and

$$(1 - \alpha_F\alpha_R)I_{CS}[\exp(qV_{CB}/kT) - 1] = \alpha_F I_E + I_C. \tag{735.6}$$

In the two-diode model, they simplify to

$$(1 - \alpha_F\alpha_R)I_{ED} = I_E + \alpha_R I_C \tag{735.5A}$$

and

$$(1 - \alpha_F\alpha_R)I_{CD} = \alpha_F I_E + I_C. \tag{735.6A}$$

These can be rearranged to give the terminal current as the sum of the current due to the terminal voltage and the transfer of a fraction of the current from the other terminal, known as the **hybrid two-port equations**,

$$I_E = (1 - \alpha_F\alpha_R)I_{ES}[\exp(qV_{EB}/kT) - 1] - \alpha_R I_C \tag{735.7}$$

and

$$I_C = (1 - \alpha_F\alpha_R)I_{CS}[\exp(qV_{CB}/kT) - 1] - \alpha_F I_E. \tag{735.8}$$

They are a mixed pair. The first is one of the two **reverse hybrid** or **g-parameter two-port equations** while the second is one of the two **forward hybrid** or **h-parameter two-port equations**.

The above can also be rearranged to give the terminal currents as a function of the current and voltage of the other terminal, known as the **transmission two-port equations**

$$I_E = - \alpha_F^{-1}I_C + [(1-\alpha_F\alpha_R)/\alpha_F]I_{CS}[\exp(qV_{CB}/kT) - 1] \tag{735.9}$$

and

$$I_C = - \alpha_R^{-1}I_E + [(1-\alpha_F\alpha_R)/\alpha_R]I_{ES}[\exp(qV_{EB}/kT) - 1]. \tag{735.10}$$

These again are a mixed pair. The first is one of the two **forward transmission two-port equations** while the second is one of the two **reverse transmission two-port equations**.

Note that these equations are all linear between the variables I_C, I_E, I_{CD} and I_{ED} where $I_{CD}=I_{CS}[\exp(qV_{CB}/kT)-1]$ and $I_{ED}=I_{ES}[\exp(qV_{EB}/kT)-1]$ if the current gain, α_F and α_R, are independent of the currents and voltages. If I_{CS} and I_{ES} are also constant, then the linearity persists between the more elementary set of variables, I_C, I_E, $\exp(qV_{CB}/kT)$ and $\exp(qV_{EB}/kT)$ or between current and charge since the charge is proportional to the Boltzmann factor, $\exp(qV/kT)$. We previously stated that this linearity is expected when we noted that the differential equations governing the minority carrier concentration and current are linear.

General Common-Emitter Two-Port Network Equations

Similar sets of two-port network equation pairs for the four variables (I_B, I_C, V_{BE}, and V_{CE}) of the common-emitter configurations can be derived from the original common-base EM equations using the Kirchoff's laws

$$I_E = -(I_B + I_C) \tag{735.11}$$

and

$$V_{CB} = V_{CE} - V_{BE} = V_{CE} + V_{EB}. \tag{735.12}$$

The algebra involves only a few steps and is obvious. The common-emitter equations with the input and output voltages, V_{BE} and V_{CE}, as the independent variables, give the **short-circuit y-parameter two-port equations** which are then

$$I_B = -(1-\alpha_F)I_{ES}\{\exp(-qV_{BE}/kT)-1\} -(1-\alpha_R)I_{CS}\{\exp[q(V_{CE}-V_{BE})/kT]-1\} \tag{735.13}$$

$$I_C = - \quad \alpha_F I_{ES}\{\exp(-qV_{BE}/kT)-1\} + \quad I_{CS}\{\exp[q(V_{CE}-V_{BE})/kT]-1\} \tag{735.14}$$

Inverting the above pair to make the input and output currents, I_B and I_C, as the independent variables, the **open-circuit z-parameter two-port equations** are

$$(1-\alpha_F\alpha_R)I_{ES}\{\exp[-q(V_{BE})/kT]-1\} = - \quad I_B - (1-\alpha_R)I_C \tag{735.15}$$

and

$$(1-\alpha_F\alpha_R)I_{CS}\{\exp[q(V_{CE}-V_{BE})/kT]-1\} = -\alpha_F I_B + (1-\alpha_F)I_C. \tag{735.16}$$

These can be rearranged to give the terminal current as the sum of the current due to the terminal voltage (true only for the input or I_B) and the transfer of a fraction of the current from the other terminal, known as the **hybrid-parameter two-port equations**,

$$I_B = -(1-\alpha_F\alpha_R)I_{ES}\{\exp(-qV_{BE}/kT)-1\} - (1-\alpha_R)I_C \tag{735.17}$$

and

$$I_C = \frac{(1-\alpha_F\alpha_R)}{(1-\alpha_F\quad)} I_{CS}\{\exp[q(V_{CE}-V_{BE})/kT] - 1\} + \frac{\alpha_F}{1-\alpha_F} I_B. \tag{735.18}$$

The first equation above belongs to the **reverse hybrid** while the second belongs to the **forward hybrid**.

These can further be rearranged to give the terminal currents as a function of the current and voltage of the other terminal (true only for the output or I_C), known as the **transmission two-port equations**,

$$I_B = [(1-\alpha_F)/\alpha_F]I_C-[(1-\alpha_F\alpha_R)/\alpha_F]I_{CS}\{\exp[q(V_{CE}-V_{BE})/kT]-1\} \tag{735.19}$$

and

$$I_C = -[1/(1-\alpha_R)]I_B-[(1-\alpha_F\alpha_R)/(1-\alpha_R)]I_{ES}\{\exp[-q(V_{BE})/kT]-1\} \tag{735.20}$$

where the first belongs to the **reverse transmission** and the second belongs to the **forward transmission two-port equations**.

Note that (735.18) and (735.19) are no longer linear even in $\exp(qV_{BE}/kT)$ and $\exp(qV_{CE}/kT)$ due to the presence of their product from $\exp(qV_{CB}/kT) = \exp[q(V_{CE}-V_{BE})/kT] = \exp(qV_{CE}/kT)\exp(-qV_{BE}/kT)$.

Circuit Models in the Four Operation Modes

These general two-port network equations will be used to synthesize the CB and CE equivalent circuits shown in Figs.735.1 to 735.6 for both the p/n/p and n/p/n BJTs. The circuits are then simplified in the four modes of operation and discussed in the following four titled subsections.

Forward Active Mode

The equivalent circuits in the forward active mode in the CB and CE configurations are given in Figs.735.1 and 735.2 respectively. Figure 735.1(a) shows the **implicit hybrid-T model** in the CB configuration and Fig.735.2(a), the **implicit hybrid-π model** in the CE configuration, which are synthesized from (735.1)-(735.4) without making any approximations. The corresponding figures (b) are the **explicit hybrid-T** and **hybrid-π** models synthesized from equations derived from making use of some simplifying approximations which are accurate in the active region. These approximations are discussed in the following paragraphs.

In this mode, the transistor is active or turned on and passes a significant amount of current. The emitter-base junction is forward biased and the collector-base junction is reverse biased. To cover the low-current range for low power applications, we shall retain both terms in $[\exp(qV_{EB}/kT) - 1]$ by not making the often-used medium-current approximation, $V_{EB} > 4kT/q \approx 100\text{mV}$, that drops the factor 1. However, we shall make this approximation for the collector-base voltage, $V_{CB} \lesssim -4kT/q \approx -100\text{mV}$, so that the exponential term in $[\exp(qV_{CB}/kT)-1]$ can be neglected compared with 1. The dropped term gives an error smaller than $\exp(-4)=0.01832$ or 1.832% error in the collector leakage current. Using this approximation in (735.1), (735.2), and (735.8) we have respectively

$$I_E \approx I_{ES}[\exp(qV_{EB}/kT) - 1] + \alpha_R I_{CS} = I_{ED} - \alpha_R I_{CS} \qquad (735.21)$$

$$I_C \approx -\alpha_F I_{ES}[\exp(qV_{EB}/kT) - 1] - I_{CS} = -\alpha_F I_{ED} - I_{CS} \qquad (735.22)$$

and

$$I_C \approx -\alpha_F I_E - (1-\alpha_F\alpha_R)I_{CS} = -\alpha_F I_E + I_{CBO} \qquad (735.23)$$

where the collector leakage current in the common-base connection with the emitter-base open circuited or $I_E=0$ is defined by

$$I_{CBO} = -(1 - \alpha_F\alpha_R)I_{CS}. \qquad (735.24)$$

The negative sign shows that the leakage current is flowing out of the collector terminal. Equations (735.21) to (735.23) are exactly the same as (731.1) to (731.3) which were derived by intuition. However, we now have the precise definitions of the parameters I_{ED}, I_{ES}, α_R, I_{CBO}, I_{CS}, and α_F, which depend on the material and device properties given in sections 733 and 734. Their definition was unknown in section 731.

The simplified equivalent circuit in the active mode of the common-base configuration can be synthesized from (735.21) and (735.23). They are shown in Fig.735.1(b) for p/n/p and n/p/n BJTs. They are known as the explicit hybrid-T model because the output current source, $\alpha_F I_E$, depends explicitly on the input current, I_E, instead of implicitly through I_{ED} in the implicit hybrid-T model given in Fig.735.1(a).

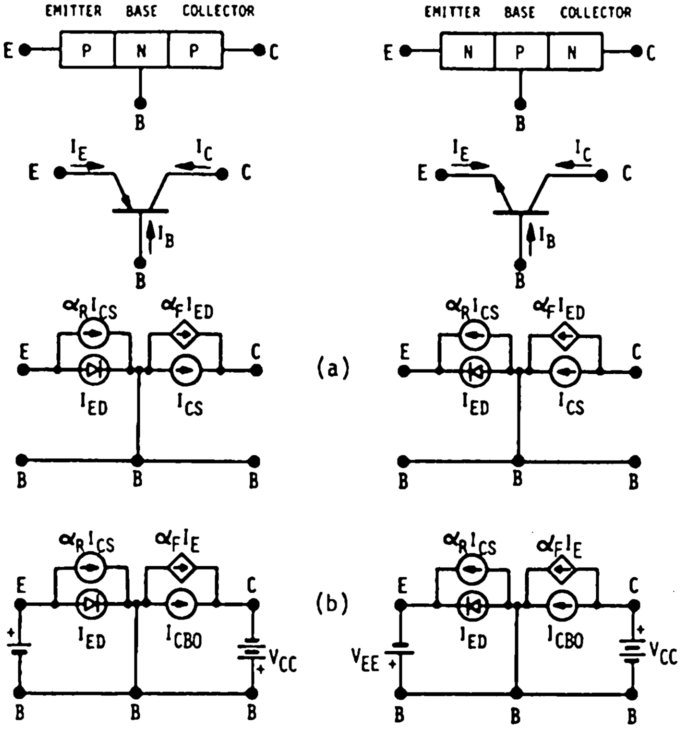


Fig.735.1 The simplified 2-port 3-terminal d.c. equivalent-circuit model of the intrinsic BJT in the **common-base forward active mode. (a)** The implicit hybrid-T model. **(b)** The explicit hybrid-T model.

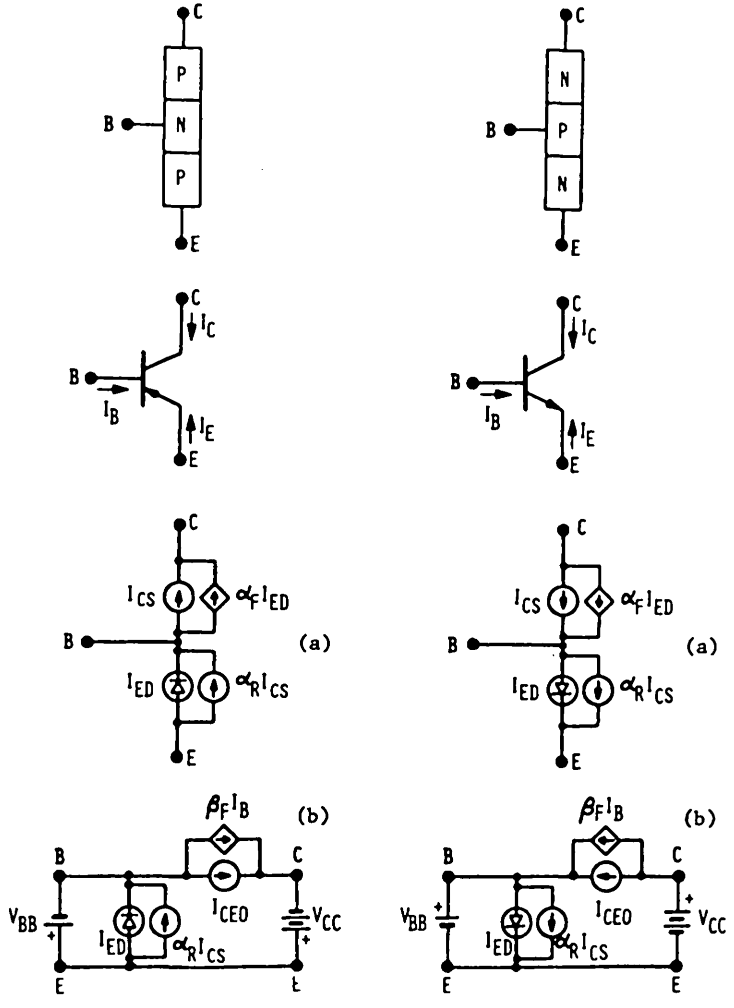

C
P
N
P
B
E
C
N
P
N
B
E
I_C
I_B
I_E
I_{CS}
$\alpha_F I_{ED}$
I_{ED}
$\alpha_R I_{CS}$
(a)
$\beta_F I_B$
(b)
V_{BB}
I_{CEO}
V_{CC}

Fig.735.2 The simplified 2-port 3-terminal d.c. equivalent-circuit model of the intrinsic BJT in the common-emitter forward active mode. (a) The implicit hybrid-π model. (b) The explicit hybrid-π model.

The equations for the forward active mode in the common-emitter configuration can be obtained similarly using again the condition that $V_{CB} \leq -4kT/q \simeq -100mV$. Again without making an approximation for V_{BE}, (735.13), (735.14) and (735.18) give

$$I_B \approx -(1-\alpha_F)I_{ES}\{\exp(-qV_{BE}/kT)-1\} - (1-\alpha_R)I_{CS} \quad (735.25)$$

$$I_C \approx - \alpha_F I_{ES}\{\exp(-qV_{BE}/kT)-1\} + I_{CS} \quad (735.26)$$

and

$$I_C \approx [\alpha_F/(1-\alpha_F)]I_B + [(1-\alpha_F\alpha_R)/(1-\alpha_F)]I_{CS}$$

$$= \beta_F I_B + I_{CEO} \quad (735.27)$$

where the forward beta or common-emitter forward gain is defined by

$$\beta_F = \alpha_F/(1-\alpha_F) \quad (735.28)$$

and

$$\alpha_F = \beta_F/(1+\beta_F). \quad (735.29)$$

The open-circuit ($I_B=0$) common-emitter collector leakage current is given by

$$I_{CEO} = +[(1-\alpha_F\alpha_R)/(1-\alpha_F)]I_{CS} = I_{CBO}/(1-\alpha_F) = (\beta_F+1)I_{CBO}. \quad (735.30)$$

The simplified and approximate explicit hybrid-π model of p/n/p and n/p/n BJTs in the active mode of the common-emitter configuration are shown in Fig.735.2(b). They can be readily synthesized from (735.25) and (735.27) without further elaboration.

There are many applications of the above results. Several will be discussed in the following numbered paragraphs.

(1) Two equivalent circuit diagrams were already drawn in the CB configuration [Fig.735.1(a) and (b)] using the I_E equation (735.21) of the emitter branch and the I_C equation (735.22) or (735.23) of the collector branch. Similarly, the I_B equation (735.25) of the base-emitter branch and (735.26) or (735.27) of the collector-emitter branch were used to draw the two equivalent circuit diagrams in the CE configuration, shown in Figs.735.2(a) and (b). Figures 735.1(b) and 735.2(b) are preferred since the emitter diode current I_{ED} of $\alpha_F I_{ED}$ in Fig.735.1(a) and 735.2(a) are internal currents which cannot be measured directly while in Fig.735.1(b), it is $\alpha_F I_E$, and in Fig.735.2(b), it is $\beta_F I_B$, and both I_E and I_B are easily measurable terminal currents.

(2) The results and the circuit models suggest simple measurements can be made to determine the EM or two-port parameters. Consider the CB configuration described by (735.21)-(735.24) and Figs.732.1 and 732.2. I_{CBO} is the measured collector current when the emitter terminal is open-circuited or $I_E=0$. I_{CS} is the measured collector current and $\alpha_R I_{CS}$ is the measured emitter current when the emitter-base terminals are short-circuited or $V_{EB}=0$ which also gives the reverse alpha, α_R. Sometimes, the emitter-base junction short-circuit condition may not be true due to a series base resistance. Knowing I_{CBO}, the forward alpha α_F can also be calculated from a series of measured I_C vs I_E as indicated by (735.23) or Fig.732.2(a). Finally, (735.22) suggests that a plot of I_C vs V_{EB} on a

semilogarithmic paper, Fig.732.2(a), will give the product $\alpha_F I_{ES}$ so that I_{ES} can be determined using the value of α_F just measured. Similarly, the parameters can also be measured in the CE configuration as indicated by (735.25)-(735.30) and Figs.732.3 and 732.4. For example, I_{CEO} is measured when the base terminal is open. β_F is measured with a set of measured base and collector currents using the measured I_{CEO}.

(3) Another interesting and useful result is the floating potential measurement across the collector-base junction previously shown in Fig.732.2(a"). Setting $I_C=0$ in (735.2) to depict floating the collector terminal, we get

$$[\exp(qV_{CB}/kT)-1]/[\exp(qV_{EB}/kT)-1] = \alpha_F I_{ES}/I_{CS} = \alpha_R. \qquad (735.31A)$$

Similarly, setting $I_E=0$ in (735.1), the floating potential across the emitter-base junction shown in Fig.732.2(b") is given by

$$[\exp(qV_{EB}/kT)-1]/[\exp(qV_{CB}/kT)-1] = \alpha_R I_{CS}/I_{ES} = \alpha_F. \qquad (735.31B)$$

Thus, the floating potential is proportional to the voltage applied to the other junction, V_{EB} or V_{CB}, when V_{EB} and V_{CB} are greater than about $4(kT/q)\simeq 100$mV. The difference between the applied and the floating potentials is proportional to the log of the alpha:

$$V_{CB}(\text{float}) = V_{EB}(\text{applied}) + (kT/q)\log_e(\alpha_R) \qquad (I_C=0) \qquad (735.32)$$

$$V_{EB}(\text{float}) = V_{CB}(\text{applied}) + (kT/q)\log_e(\alpha_F) \qquad (I_E=0) \qquad (735.33)$$

(4) One simple and exact result can be obtained from the measurements and the theory. It concerns the product $\alpha_F I_{ES}$ discussed in (2) above and it is not shown by the EM equations. From the definition of α_F and I_{ES} given in section 734, we have

$$\alpha_F I_{ES} = \alpha_B j_B A_E = \text{sech}(X_B/L_B)(qD_B P_B/L_B)\text{ctnh}(X_B/L_B)A_E \qquad (735.34A)$$

$$= (qD_B P_B/L_B)A_E/\sinh(X_B/L_B). \qquad (735.34B)$$

For a high-gain transistor, the base is thin compared with the diffusion length, $X_B << L_B$, thus, $\sinh(X_B/L_B)\simeq(X_B/L_B)$. Then, this product reduces to

$$\alpha_F I_{ES} = A_E q D_B P_B/X_B = A_E q D_B n_i^2/(N_{DD}X_B) \qquad (735.35)$$

which shows that it is indeed a constant, not affected by recombination-generation currents in the junction space-charge layers. Thus, the plot of $\log(I_C)$ vs V_{EB} [Fig.732.2(a)] will give a true straight line whose slope measures the total impurity concentration in the base layer, $N_{DD}X_B$. This is known as the Gummel number to be discussed in section 737. It is this true exponential dependence which is used to design logarithmic and exponential amplifiers.

Reverse Active Mode

The results of the reverse active mode in the two configurations are the same as the forward active mode with the emitter and collector subscripts exchanged.

Cutoff Mode

When both the emitter-base and the collector-base junctions are reverse biased, and $|V_{EB}|$ and $|V_{CB}|$ > $4(kT/q) \approx 100$mV, only leakage currents flow through the two diodes. The CB equivalent circuit shown in Fig.735.3(a) is obtained from (735.1). Since the leakage currents are small, the two current sources can be dropped and the three terminals (emitter, base and collector) are effectively disconnected from each other as shown in Fig.735.3(b). A similar trivial model is obtained in the CE configuration and shown in Fig.735.4(b).

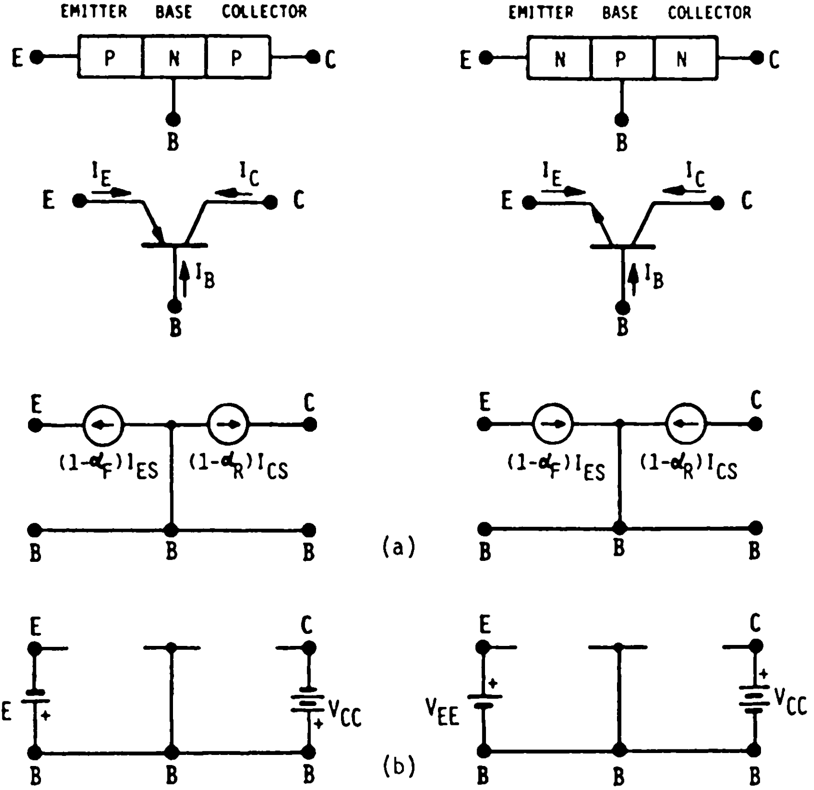


Fig.735.3 The simplified 2-port 3-terminal d.c. equivalent-circuit model of the intrinsic BJT in the common-base cutoff mode. (a) The hybrid-T model. (b) The open-circuit approximation neglecting leakage current sources.

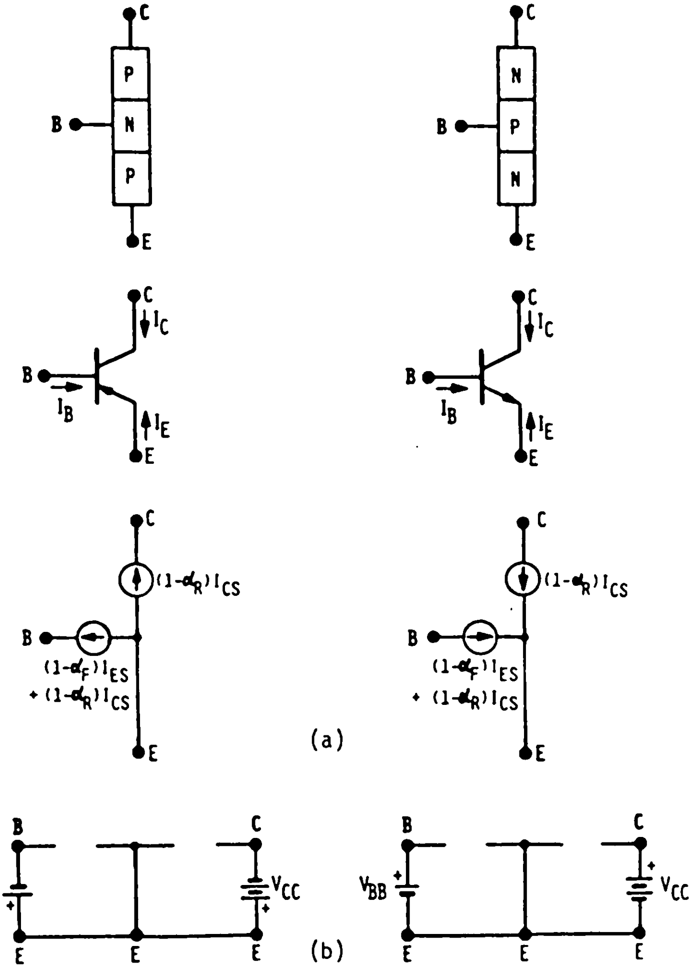


Fig.735.4 The simplified 2-port 3-terminal d.c. equivalent-circuit model of the intrinsic BJT in the common-emitter cutoff mode. (a) The hybrid-π model. (b) The open-circuit approximation neglecting leakage current sources.

Saturation Mode

In the saturation mode, both the emitter-base and collector-base junctions are forward biased. This can be readily measured in the common-base configuration by forcing a emitter current to give a forward voltage of $|V_{EB}|>4kT/q=100mV$, and measuring the collector-base voltage in the range of $|V_{CB}|>\approx 30kT/q=700mV$ as indicated by Figs.732.1(a) in the common-base configuration of a p/n/p BJT. It is more frequently measured in the common-emitter configuration by forcing both a large constant base current and constant collector current into the transistor in order to bias it into the saturation range of the characteristics. The experimental characteristics were shown in Fig.732.4(a). Using the condition of forward bias in (735.15) and (735.16) and dividing the latter by the former, the saturation voltage is then

$$V_{CE}(\mathrm{sat}) = (kT/q)\log_e\left[\frac{[-\ \alpha_F I_B + (1-\alpha_F)I_C]I_{ES}}{[-\ I_B - (1-\alpha_R)I_C]I_{CS}}\right] \tag{735.36A}$$

$$= (kT/q)\log_e\left[\frac{1 - (I_C/I_B)/\beta_F}{1/\alpha_R + (I_C/I_B)/\beta_R}\right] \tag{735.36B}$$

$$= (kT/q)\log_e\left[\frac{1 - (I_C/I_B)/\beta_F}{1 + [1+(I_C/I_B)]/\beta_R}\right]. \tag{735.36C}$$

Here we have used $\alpha_F I_{ES} = \alpha_R I_{CS}$, $\beta_F = \alpha_F/(1+\alpha_F)$, and $(1/\alpha_R) = 1 + (1/\beta_R)$. Sometimes, the term forced alpha and forced beta are used to indicate that they are the values at the forced base and collector current values since the alpha and beta are functions of current. For a nearly symmetrical transistor with high gain, and at not too high an (I_C/I_B) ratio, the terms $(I_C/I_B)/\beta_F$ and $(I_C/I_B)/\beta_R$ can be dropped in (735.36A) and

$$V_{CE}(\mathrm{sat}) \approx (kT/q)\log_e(\alpha_R) < 0. \tag{735.37}$$

In real BJT, it is usually quite asymmetrical with $\beta_R = 2$ and $\beta_F = 200$. Let $I_C/I_B = 10$, then

$$V_{CE}(\mathrm{sat}) = (kT/q)\log_e\left[\frac{1 - 10/200}{1 + (1+10)/2}\right]$$

$$= (kT/q)\log_e(0.95/6.5) = -\ 49.1\mathrm{mV}. \tag{735.38}$$

This is a rather small value indicating that the forward bias across the collector-base junction is slightly smaller than that of the emitter-base junction. In real transistors, a large collector series resistance would cause comparable voltage drop. At 10mA and 5Ω collector series resistance, the ohmic drop would be 50mV so that the total $V_{CE}(\mathrm{sat})$ is $-49.1 - 50 \approx -\ 100$ mV.

Other results in the saturation mode are summarized as follows. They can be readily derived using the procedure just described.

$$V_{EB}(\mathrm{sat}) = (kT/q)\log_e\left[\frac{\alpha_R I_B - (1-\alpha_R)I_E}{I_{EBO}}\right] \tag{735.39A}$$

$$= (kT/q)\log_e\left[\frac{I_B + (1-\alpha_R)I_C}{I_{EBO}}\right]. \tag{735.39B}$$

To get V_{CB}(sat), we can simply interchange the subscripts E and C in V_{CB}(sat).

$$V_{CB}(\mathrm{sat}) = (kT/q)\log_e\left[\frac{\alpha_F I_B - (1-\alpha_F)I_C}{I_{CBO}}\right] \tag{735.40A}$$

$$= (kT/q)\log_e\left[\frac{I_B - (1-\alpha_F)I_E}{I_{CBO}}\right]. \tag{735.40B}$$

Two asymptotic forms of the V_{CE}(sat) can be obtained from (735.36A)-(735.36C) which are particularly useful for BJT design.

If (i) $\beta_F >> (I_C/I_B)$, which is commonly attainable since β_F can be 100-200 and the normal test condition is $I_C/I_B = 10$; and also if (ii) $(I_C/I_B) << (\beta_R/\alpha_R) = 1/(1-\alpha_R) \approx \beta_R$ which can be attained in a symmetrical or not highly asymmetrical transistor (i.e. $\alpha_R \approx \alpha_F \approx 1$) and both β_F and β_R are high, then

$$V_{CE}(\mathrm{sat}) \approx (kT/q)\log_e\left[\frac{1 - (I_C/\beta_F I_B)}{\alpha_R^{-1} + (I_C/\beta_R I_B)}\right] \tag{735.41A}$$

$$\approx (kT/q)\log_e \alpha_R \approx -(kT/q)(1-\alpha_R) = \mathrm{small} \tag{735.41B}$$

This shows how to design BJT with small V_{CE}(sat), a particularly useful rule in power BJTs. An implicit assumption is that β_F and β_R are large in the saturation mode and not just in the active mode.

If both α_R and β_R are small, $(I_C/I_B) >> (\beta_R/\alpha_R) = 1/(1-\alpha_R) \approx 1$, and also if $\beta_F >> (I_C/I_B)$ as in a normal transistor, then

$$V_{CE}(\mathrm{sat}) \approx (kT/q)\log_e\left[\frac{1 - (I_C/\beta_F I_B)}{(I_C/\beta_R I_B)}\right] \tag{735.42A}$$

$$\approx (kT/q)\log_e[\beta_R I_B/I_C] \tag{735.42B}$$

which gives a direct measure of the reverse beta. Another key point is that the V_{CE}(sat) is not affected by the base resistance, $r_{b'}$, since a constant I_B drive is used to measure the V_{CE}(sat).

The CB and CE equivalent circuit models in the saturation range are rather simple. They are T-networks, each containing two voltage sources representable by batteries and three resistances. The three resistances consist of the base-spreading, series, and bulk resistances in series with contact resistances and saturation resistances. The T-equivalent circuit for the common-base configuration is given in Figs.735.5(b). V_{EB}(sat) in this figure is given by (735.39A) and (735.39B), and V_{CB}(sat) is given by (735.40A) and (735.40B). The T-equivalent circuit for the common-emitter configuration is given in Figs.736.6(c). V_{EB}(sat) in this CE configuration is the same as that in the CB configuration given by (735.39A) and (735.39B). V_{CB}(sat) is given by (735.36A), (735.36B), and (735.36C), and is about 50mV as estimated by (735.38). The collector bulk and contact series resistances give another 50mV drop making the total about 100mV as indicated after (735.38).

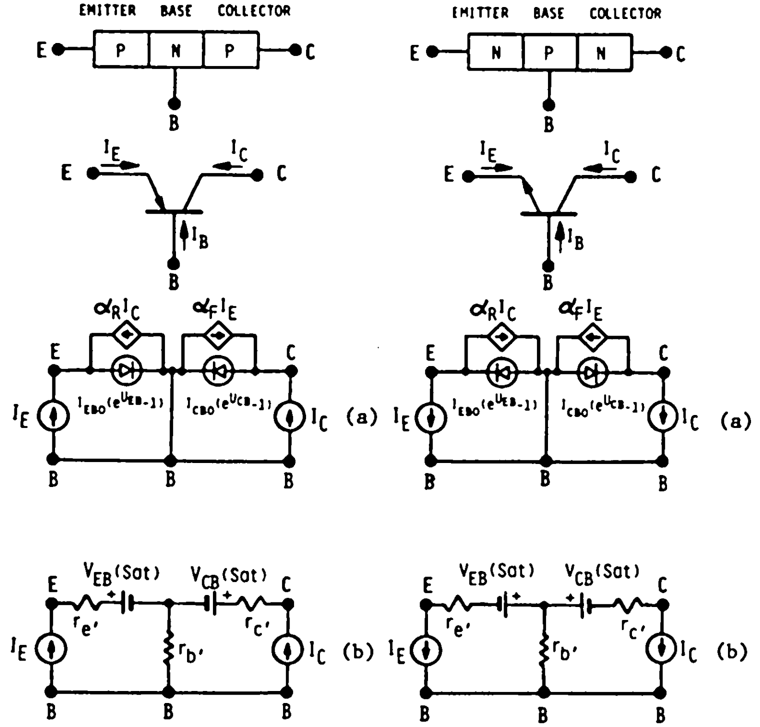

EMITTER
BASE
COLLECTOR
P
N
P
N
P
N
E
C
B
I_E
I_C
I_B
$\alpha_R I_C$
$\alpha_F I_E$
$I_{EBO}(e^{U_{EB}}-1)$
$I_{CBO}(e^{U_{CB}}-1)$
(a)
V_{EB}(Sat)
V_{CB}(Sat)
$r_{e'}$
$r_{b'}$
$r_{c'}$
(b)

Fig.735.5 The simplified 2-port 3-terminal d.c. equivalent-circuit model of the intrinsic BJT in the **common-base saturation mode**. (a) The open-circuit-parameter T model. (b) The variable battery model.

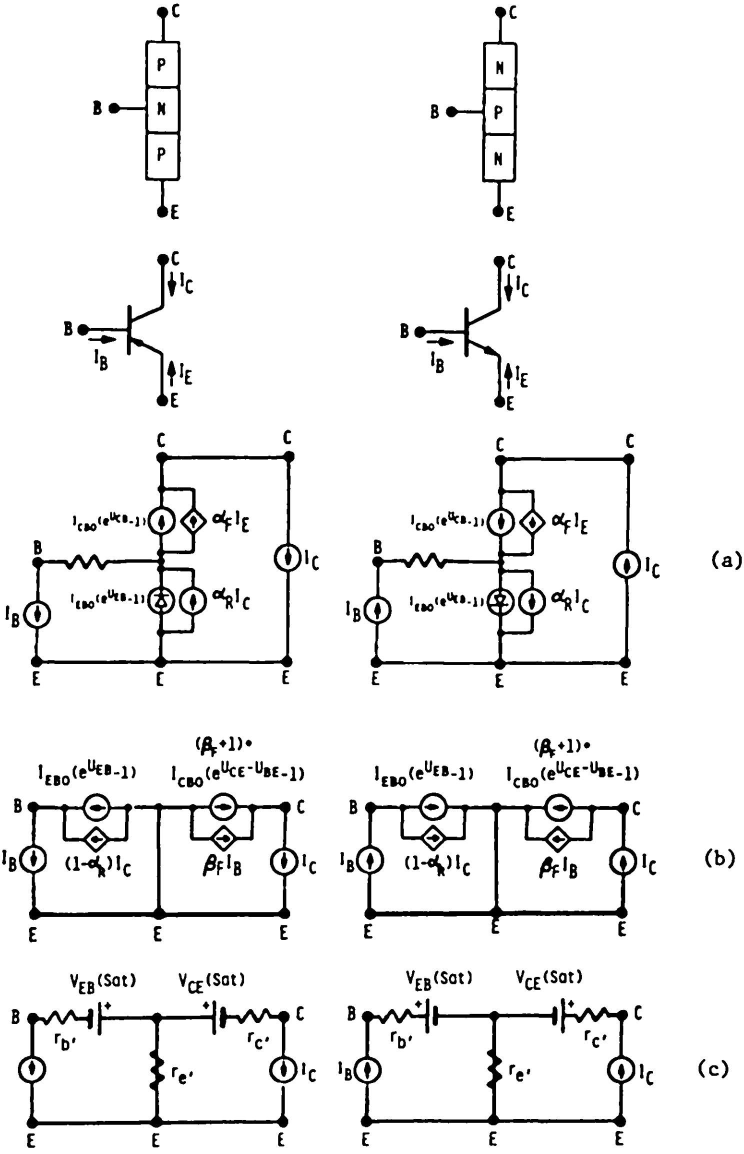

C
P
N
P
B
E
C
N
P
N
B
E
I_C
I_B
I_E
$I_{CBO}(e^{U_{CB}}-1)$
$\alpha_F I_E$
$I_{EBO}(e^{U_{EB}}-1)$
$\alpha_R I_C$
I_B
I_C
(a)
$(\beta_F+1)\cdot$
$I_{EBO}(e^{U_{EB}}-1)$
$I_{CBO}(e^{U_{CE}-U_{BE}}-1)$
$(1-\alpha_R)I_C$
$\beta_F I_B$
(b)
$V_{EB}(Sat)$
$V_{CE}(Sat)$
$r_{b'}$
$r_{e'}$
$r_{c'}$
(c)

Fig.735.6 The simplified 2-port 3-terminal d.c. equivalent-circuit model of the intrinsic BJT in the **common-emitter saturation mode.** **(a)** The rotated CB open-circuit-parameter T model. **(b)** The nonlinear hybrid-π model. **(c)** The variable battery model.

736 Lumped D.C. Models of Realistic Multi-Dimensional BJT

To analyze and design realistic BJTs, parasitic elements due to 2-d (two-dimensional) and even 3-d effects must be added to the original-extended Ebers-Moll or SNS d.c. intrinsic BJT models. The multi-dimensional effects will be analyzed by applying the preceding 1-d results to each region of a partitioned BJT.

Non-overlap Collector Diode and Base Spreading Resistance

The important parasitic elements are the **non-overlap collector-base** n/p **junction diode** briefly discussed in the previous sections and the **lateral base resistance** due to the thin base necessary to give high speed and frequency. Other parasitics include the resistance of the semiconductor body and diffused regions, and the contact resistances. They can all be included in the two-port three-terminal intrinsic transistor models as one-lump elements as indicated in Fig.736.1(b): a p/n junction diode connected in parallel with the collector-base junction, and a resistance in series with terminal of the base, $r_{b'}$, collector, $r_{c'}$, and emitter, $r_{e'}$. This one-lump model is a good approximation in a well-designed BJT operated at low currents. However, the constant $r_{b'}$ approximation breakdowns easily even in well-designed BJT when the base current increases to give a voltage drop greater than kT/q across $r_{b'}$ because the emitter-base diode current is highly nonlinear (exponentially varying with the local emitter-base junction voltage). Thus, the edge of the emitter-base junction has the full bias V_{EB} and much higher emitter and collector current densities while the central part has lower bias $V_{EB}-I_Br_{b'}$ and much lower current density. This current crowding also invalidates the one-lump approximation of $r_{e'}$ and $r_{c'}$. Figure 736.1(a) illustrates the parasitic and 2-d effects and Fig.736.1(c) the distributed model to be discussed later.

The terminal current-voltage characteristics of the real BJT can be readily written down using the extended Ebers-Moll equations just developed for the intrinsic BJT and the diode equation given in chapter 5 for non-overlap collector-base diode. Let the emitter-base junction area be A_E and the collector-base junction area be A_C, then by inspection of Figs.736.1(a) and (b), we have

$$I_E = J_EA_E = + \quad I_{ES}[\exp(qV_{EB}'/kT)-1] - \alpha_RI_{CS}[\exp(qV_{CB}'/kT)-1] \qquad (736.1)$$

$$\begin{aligned} I_C &= J_CA_E + J_{PN}(A_C-A_E) \\ &= - \alpha_FI_{ES}[\exp(qV_{EB}'/kT)-1] + \quad I_{CS}[\exp(qV_{CB}'/kT)-1] \\ &\qquad + \quad I_{CS}'[\exp(qV_{CB}'/kT)-1]. \end{aligned} \qquad (736.2)$$

The internal voltages are labeled with a prime, V_{EB}' and V_{CB}', and given by

$$V_{EB}' = V_{EB} - I_Br_{b'} - I_Er_{e'} = V_{EB} - (I_E+I_C)r_{b'} - I_Er_{e'} \qquad (736.3A)$$

and

$$V_{CB}' = V_{CB} - I_Br_{b'} - I_Cr_{c'} = V_{CB} - (I_E+I_C)r_b' - I_Cr_{c'}. \qquad (736.3B)$$

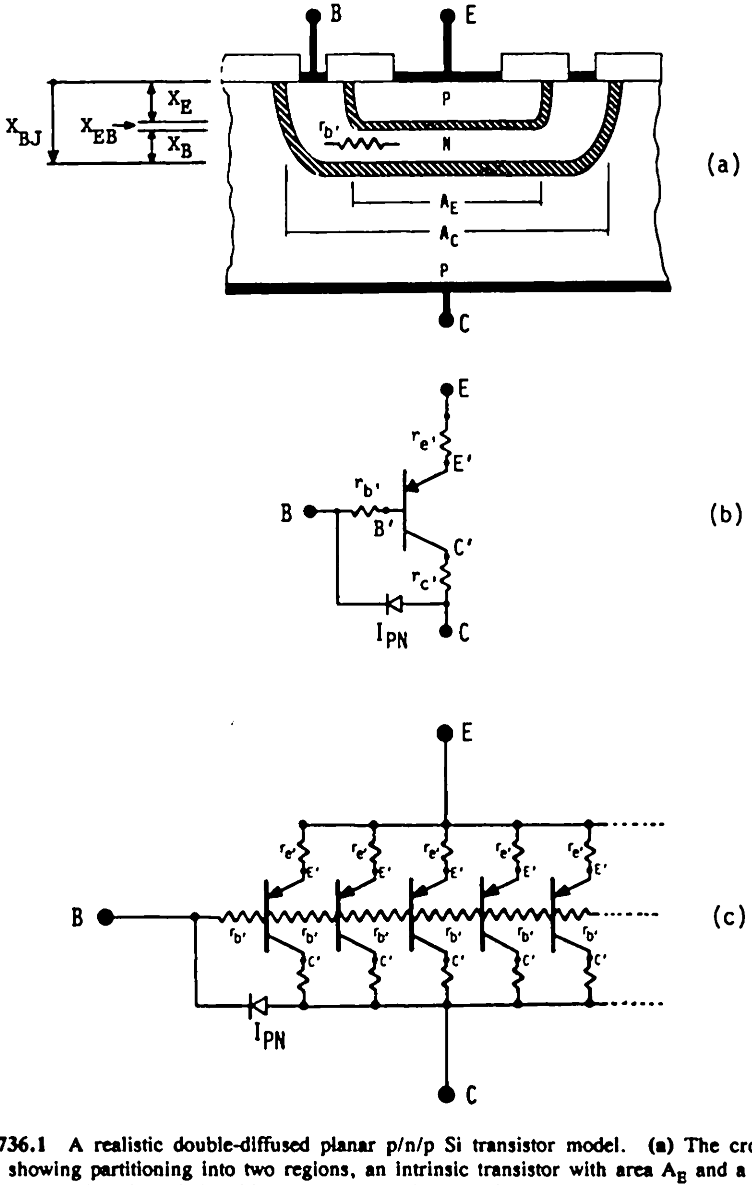


Fig.736.1 A realistic double-diffused planar p/n/p Si transistor model. **(a)** The cross-sectional view showing partitioning into two regions, an intrinsic transistor with area A_E and a non-overlap collector n/p junction diode with area A_C-A_E. **(b)** The circuit symbols and circuit model of the transistor with a lumped lateral base resistance, emitter and collector contact and bulk resistances, and a non-overlap collector n/p diode. **(c)** The distributed base resistance model which also distributes the emitter and collector resistances.

The current through the non-overlap collector-base junction diode, represented by the term with J_{PN} or I_{CS}' in (736.2), is obtained from the collector-base junction current equation (733.20) with V_{EB} set to zero. This is used instead of the thick-base diode equation given in chapter 5 since the base of this non-overlap collector diode is thin and it has an ohmic contact at the surface of the thin base layer. The current coefficients j_B must also be modified since the n-layer of the non-overlap collector n/p junction diode, $X_{BJ}=X_E+X_{EB}+X_B$ in Fig.736.1(a), is thicker than the base layer under the emitter of the intrinsic transistor, X_B. Thus,

$$I_{CS}' = (A_C - A_E)(j_{RC} + j_C + j_B') \qquad (736.4)$$

where

$$j_B' = (qD_BP_B/L_B)\mathrm{ctnh}[(X_{BJ})/L_B]$$

$$= (qD_BP_B/L_B)\mathrm{ctnh}[(X_E+X_{EB}+X_B)/L_B] \qquad (736.5)$$

while for the intrinsic transistor, we had in (733.19C)

$$j_B = (qD_BP_B/L_B)\mathrm{ctnh}[(X_B)/L_B].$$

From the above analysis, it is evident that the realistic BJT transistor model is easily obtained by simply combining the extended Ebers-Moll equations for the intrinsic transistor and the current-voltage characteristics of the parasitic elements, in this case, a series base resistance, a series collector bulk and contact resistance, a series emitter bulk and contact resistance, and a non-overlap collector p/n junction diode. **This simple procedure can be used for even more complex BJT geometries by judicially partitioning the transistor into several regions and applying the simple and valid 1-d transistor, diode, or resistance equations to each region.** It is superior than a black-box approach since (i) the model parameters of the regional elements in this model are all derived based on diode and transistor physics and (ii) the underlying assumptions are concisely and quantitatively known. What is neglected in the 1-d partitioning model is the effect from fringing or non-parallel electric field lines and non-parallel current flow lines since the 1-d model assumes all the lines are parallel. But, these convergent-divergent or nonparallel current flow and electric field lines can be approximated by adjusting the areas of each region. This adjustment would give a sufficiently accurate model in most applications while maintaining the model's algebraic simplicity. Simplicity is crucial in computer-aided-design (CAD) applications of the BJT models to design VLSI chips. This is because a VLSI chip contains so many (10^4 to over 10^6) transistors that accurate and fast-convergent numerical results can be obtained within a reasonable CPU iteration time only if each of the ten-thousand to more than one-million transistors is represented by a simple few-parameter device model.

Numerical solutions of the real-transistor equations are more difficult and tedious to obtain than the ideal and extended Ebers-Moll equations since they are transcendental due to the presence of the base resistance which makes $V_{EB}' \neq V_{EB}$ and $V_{CB}' \neq V_{CB}$. But this is only one numerical solution cycle for one transistor. It is still much simpler and much faster to obtain than to seek numerical solutions from two- or three-dimensional (2-D or 3-D) nonlinear coupled-differential equations for each transistor since the multi-dimensional solutions would require the use of a large matrix of points (x,y,z) for each transistor, such as a 100x100 2-D or 100x100x100 3-D arrays of grid points (nodes), to get accurate current-voltage solutions at each point and at terminal nodes.

For the 1-lump model of Fig.736.1(b), a numerical iteration sequence can readily be started by assuming $r_{b'} = r_{c'} = r_{e'} = 0$ as the zeroth approximation and then the values of V_{EB} and V_{CB} can be corrected by computing the voltage drop through $r_{b'}$, $r_{c'}$ and $r_{e'}$. For well-designed high-performance Si BJTs, $r_{b'}$ is the most important among the three resistors since high-frequency and high-speed require thin base resulting in large $r_{b'}$. Its importance can be demonstrated readily by a back-of-the-envelope estimate. Taking the previous numerical example: $N_B = N_N \approx N_{DD} = 10^{17}\text{cm}^{-3}$ and $X_B = 10^{-4}\text{cm}$ for the n-base, (Width)x(Length) = $(W_E x L_E)$ = $2\mu\text{m} x 5\mu\text{m}$ for the two emitter stripes of Fig.720.2(b), and electron mobility of $1000\text{cm}^2/\text{V-s}$ (a little high) in the n-type base layer, then the base resistance is

$$\begin{aligned} r_{b'} &= (\text{Resistivity})\text{x}(\text{Length}) \div (\text{Area}) = (1/q\mu_n N_B)(W_E/2)/(X_B \text{x} L_E) \\ &= (1.6\text{x}10^{-19}\text{x}1000\text{x}10^{17})^{-1}(10^{-4})/(10^{-4}\text{x}5\text{x}10^{-4}) \\ &= 125 \text{ ohms.} \end{aligned} \tag{736.6}$$

In the preceding section, we computed a low injection limit of 0.016 mA for the emitter and collector current at $V_{EB} = 822\text{mV}$. Thus, the voltage drop across the lateral base resistance is only $125\text{x}0.016\text{x}10^{-3} = 0.002\text{V} = 2\text{mV}$ if all the emitter current flows out of the base. This voltage drop is small and can be neglected as a first approximation in a realistic transistor operated at low injection levels. The result does show two important design considerations for minimizing base resistance. If the base contact were at the length ends of the emitter stripes, then we would have a longer (5/2=2.5 times longer) and narrower (2/5=0.4 times narrower) base resistor or 10 times higher base resistance which would increase the base voltage drop to 0.02V=20mV. This can no longer be neglected since it is comparable to kT/q=25mV. The second design consideration concerns the base layer thickness. To increase the frequency response (sections 74n) or the switching speed (sections 75n), the base layer must be made thinner, for example, down to $0.2\mu\text{m}$ or $0.1\mu\text{m}$ from $1\mu\text{m}$ just assumed. Then the base resistance would be higher and the voltage drop across the base resistance would increase further and become a dominant limitation on achieving higher frequency response and speed.

From an inspection of the cross-sectional view of the double-diffused planar p/n/p BJT shown in Fig.736.1(a), one can readily conclude that the series base resistance is distributed. This can be represented by a distributed BJT model shown in Fig.736.1(c). If the voltage drop along the distributed base resistance is significant, then the intrinsic transistor must also be treated as a distributed transistor. The one-dimensional transistor model still gives an accurate representation of the individual distributed transistor but the numerical calculation for the external current-voltage characteristics becomes more tedious. Analytical solution is not possible but it is straightforward to solve as a circuit problem and get numerical answers using a high speed computer or even a hand-held programmable calculator.

737 Material and Structural Dependences of D.C. Two-Port Parameters of BJT

The four two-port short-circuit parameters (I_{ES}, I_{CS}, α_F, and α_R) in the d.c. BJT equations were given by (734.5) to (734.10B) with the definitions given by (733.21A) to (733.21F). Treating them as constant parameters, two-port nonlinear d.c. network models and multi-dimensional lumped models were described and analyzed in the preceding two sections. However, even in the ideal one-dimensional intrinsic BJT, these four parameters are dependent on material properties and device geometries which will be discussed in this section. They also vary with the biasing conditions (V_{CB}, V_{EB}, V_{CB}, I_C, I_E, I_B) which increase the nonlinearity. The bias dependences are discussed and analyzed in the following two sections, 738 and 739.

These dependencies must be known in order to design and characterize BJTs. They will be analyzed systematically in the sequence of increasing current density and then increasing voltage because the effects from each dependency interact. For example, the Kirk effect reduces the Early effect. Thus, this section begins the analyses of the material and structural effects on BJT's d.c. characteristics because they appear at the lowest bias voltages and currents, even at zero bias. They originate in the three quasi-neutral layers: the emitter, base and collector layers. Bias dependences, described in the next two sections, originate in the two space-charge layers whose thickness variation with bias also affects the parameters of the three quasi-neutral layers, especially the base layer. The bias dependencies give rise to the interaction among these material and structural effects.

The material and structure dependences of the four 2-port parameters in the EM equation (I_{ES}, I_{CS}, α_F and α_R) are contained in the five current coefficients (j_E, j_{EB}, j_B, j_{CB}, j_C) in the extended EM equations and the hyperbolic functions of the normalized thickness of the emitter, base and collector layers. The normalization parameter is the minority carrier diffusion length in the three quasi-neutral layers. These expressions will now be simplified to isolate the factors which control the minority carrier diffusion-recombination currents in the three quasi-neutral layers, especially the quasi-neutral emitter and base layers. The simplified results also suggest experimental conditions to measure the material parameters of a BJT.

Diffusion-Drift Transport Time in the Quasi-Neutral Base Layer

The minority carrier (hole in the p/n/p BJT) diffusion current in the base layer is given by the coefficient $\alpha_B j_B$ as indicated in (733.20). The base transport factor, α_B, is the fraction of the holes which have escaped recombination with majority carriers (electrons in the n-base) and safely reached the collector-base junction by diffusion, and also by drift if there is a electric field in the quasi-neutral base due to composition variations, $N_{AA}(x)$-$N_{DD}(x)$ and Ge_xSi_{1-x} where the x-atomic-fraction varies with x-axis. Thus, this fraction is given by the ratio $J_P(x=X_B,V_{EB})/J_P(x=0_B,V_{EB})$. Using (733.6A) and (733.6B),

$$\alpha_B = J_P(X_B,V_{EB})/J_P(0_B,V_{EB}) = \mathrm{csch}(X_B/L_p)/\mathrm{ctnh}(X_B/L_p)$$

$$= 1/\mathrm{csch}(X_B/L_p) = \mathrm{sech}(X_B/L_B) \qquad (737.1)$$

where L_p is replaced by the symbol L_B for the **minority carrier diffusion length in the base**. This is cited as a definition in (733.21D) which has now gained physical significance.

For a high gain transistor, we must have a minimum recombination loss in the base layer. This means that the base layer thickness must be much smaller than the minority carrier diffusion length, $X_B << L_B$. Equation (737.1) is then simplified using the Taylor series expansion of the hyperbolic secant [737.1],

$$\alpha_B = \mathrm{sech}(X_B/L_B) \qquad (737.2)$$

$$= 1 - (X_B/L_B)^2/2! + 5(X_B/L_B)^4/4! - .. \simeq 1 - (X_B/L_B)^2/2 \qquad (737.2A)$$

$$= 1 - X_B{}^2/(2D_B\tau_B) = 1 - (X_B{}^2/2D_B)/\tau_B \equiv 1 - t_B/\tau_B. \qquad (737.2B)$$

The characteristic time, $t_B \equiv X_B{}^2/2D_B$, is known as the **minority carrier base transport time**. It is modified by a multiplier η if there is a drift field in the base as indicated by (734.10A) and (734.10B).

Gummel Number in the Quasi-Neutral Base Layer

The base current coefficient, j_B, can also be simplified for thin base or high gain transistors. From the definition given by (733.21C) and using the Taylor series expansion for the hyperbolic cotangent [737.1], we have

$$j_B = (qD_BP_B/L_B)\mathrm{ctnh}(X_B/L_B)$$

$$\simeq (qD_BP_B/L_B)[1/(X_B/L_B)]$$

$$= (qD_BP_B/X_B) = qD_Bn_i{}^2/(N_{DD}X_B). \qquad (737.3)$$

The product ($N_{DD}X_B$) is known as the **Gummel Number**. Gummel first arrived at this relationship in 1961 [737.2]. Extensive experimental data on a variety of Si BJTs were compared with the theory independently by Sah in 1961-1962 [737.3]. Gummel then showed in 1970 its importance as a fundamental transistor parameter in BJT modeling and design [737.4] which coined the term, Gummel number. For a nonuniformly doped or diffused base BJT where the majority carrier (electron in p/n/p) concentration, mobilities, and energy gap (due to high-doping effect) can vary with thickness position, j_B is

$$j_B = \frac{q}{\int_{0B}^{X_B} [N_{DD}(x)/n_o^2(x)D_B(x)]dx} . \qquad (737.4)$$

Thus, N_{DD} in the Gummel number, ($N_{DD}X_B$), is now precisely defined. It is the average of $N_{DD}(x)$ or more precisely, $N(x)$, over the thickness of the base layer. For the numerical example we have used, the transistor has a one-micron base layer, $X_B = 1\mu m$, with a based doping of $10^{17}cm^{-3}$. The Gummel number for this transistor is then

$$N_{DD}X_B = 10^{17}x10^{-4} = 10^{13}\ cm^{-2}. \qquad (737.5A)$$

It is the areal density of the majority carriers stored in the base layer. Using $n_i(297K) = 10^{10}cm^{-3}$ for Si, the coefficient of the diffusion current is

$$\begin{aligned} j_B &= 1.6x10^{-19}x10^{20}/(10^{17}x10^{-4}) \\ &= 1.6x10^{-12}\ A/cm^2 = 1.6\ pA/cm^2. \qquad (737.5B) \end{aligned}$$

[737.1] See H.B.Dwight, **Tables of Integrals and Other Mathematical Data**, MacMillian Company, 1961, formulae 657.5 on page 154, or another mathematical table or handbook.

[737.2] H. K. Gummel, "Measurement of the number of impurities in the base layer of a transistor," Proc. IRE, 49, p.834, 1961.

[737.3] C. T. Sah, "Effect of surface recombination and channel on p/n junction and transistor characteristics," IRE Transaction on Electron Devices, ED-9(1), 94-108, Jan. 1962.

[737.4] H. K. Gummel, "A charge control relationship for bipolar transistors," Bell System Tech. J. 49(1), 827-851, Jan.-Feb. 1970.

The Gummel number and j_B are very important transistor design parameters since they determine the minority carrier current density injected into the quasi-neutral base layer from the emitter layer. It is the maximum current that can be collected by the collector-base junction. They can be measured very accurately. The plot of $\log_e I_C$ vs V_{EB} will be a straight line when $V_{EB}>4kT/q$ which was observed in the data sheet, Figs.732.2(a) and (b) for the n/p/n BJT example. The intercept of this straight line with the $V_{EB}=0$ axis will then give $j_B A_E$. This exponential relationship can be derived using one of the short-circuit-parameter BJT transistor equations, (733.20). It is worked over to simplify as follows:

$$J_C = +\ (j_C+j_{CB}+j_B)[\exp(qV_{CB}/kT)-1] - \alpha_B j_B[\exp(qV_{EB}/kT)-1] \qquad (733.20)$$

$$\approx -\ (j_C+j_{CB}) \quad -\ (1-\alpha_B)j_B \quad -\ \alpha_B j_B\cdot\exp(qV_{EB}/kT). \qquad (737.6)$$

The approximation, (737.6), is obtained when the collector-base junction is reverse biased or $V_{CB}<-2kT/q=-0.05V$. When the collector-base junction is short-circuited or $V_{CB}=0$, instead of (737.6), (733.20) reduces to

$$J_C = -\ \alpha_B j_B[\exp(qV_{EB}/kT) - 1]. \qquad (737.7)$$

The total collector current is then obtained by multiplying the above result by the junction area. We assume that the emitter-base and collector-base junction areas are not equal, then for a reverse biased collector,

$$I_C = -\ (j_{CB}+j_C)A_C - (1-\alpha_B)j_B A_E \qquad -\ \alpha_B j_B A_E\cdot\exp(qV_{EB}/kT) \qquad (737.8)$$

$$\approx -\ (j_{CB}+j_C)A_C - (qX_B n_i^2/2\tau_B N_{DD})A_E - \alpha_B j_B A_E\cdot\exp(qV_{EB}/kT) \qquad (737.8A)$$

The second term above is reduced using $1-\alpha_B \approx X_B^2/2D_B\tau_B$ from (737.2) and $j_B \approx qD_B n_i^2/(N_{DD}X_B)$ from (737.3). In the third term, $\alpha_B \approx 1$ can be assumed which is valid in a high gain transistor. Thus, for the short-circuited collector-base ($V_{CB}=0$), we have

$$I_C = -\ \alpha_B j_B A_E\cdot[\exp(qV_{EB}/kT) - 1] \qquad (737.9)$$

$$\approx -\ \alpha_B j_B A_E\cdot\exp(qV_{EB}/kT) + j_B A_E. \qquad (737.9A)$$

Note, the **emitter-base junction area**, A_E, is used in (737.9) to calculate the total collector current and not the collector-base junction area, since this collector current originates from the minority carriers (holes in p/n/p) injected by the emitter-base junction. Even if the transistor is not well designed so that α_B is significantly smaller than unity or the emitter-base junction area is larger than the collector-base junction area, the exponential dependence of I_C on V_{EB} still holds as (737.9) shows. This exponential dependence is also independent of the geometrical details and variations. For example, if the emitter-base and collector-base junction planes are not parallel or the base layer thickness varies, this exponential emitter-base voltage dependence of the collector current is still valid. The measured parameter, $j_B A_E$, would then be a geometrical average of $\alpha_B j_B$ over the collector-base junction area. The integral to account for the non-constant base thickness can be written down at once from (737.3) and (737.4). Furthermore, this exponential dependence is independent of recombination losses in the emitter-base junction space-charge layer and on the surfaces. The reason is that the collector current measures only the minority carriers that reach the collector and the minority carrier density is given by the Boltzmann factor of the emitter-base voltage, $\exp(qV_{EB}/kT)$. This is one of the most important consequences of the minority carrier injection theory formulated and recognized by Shockley in 1947 and published in 1949.

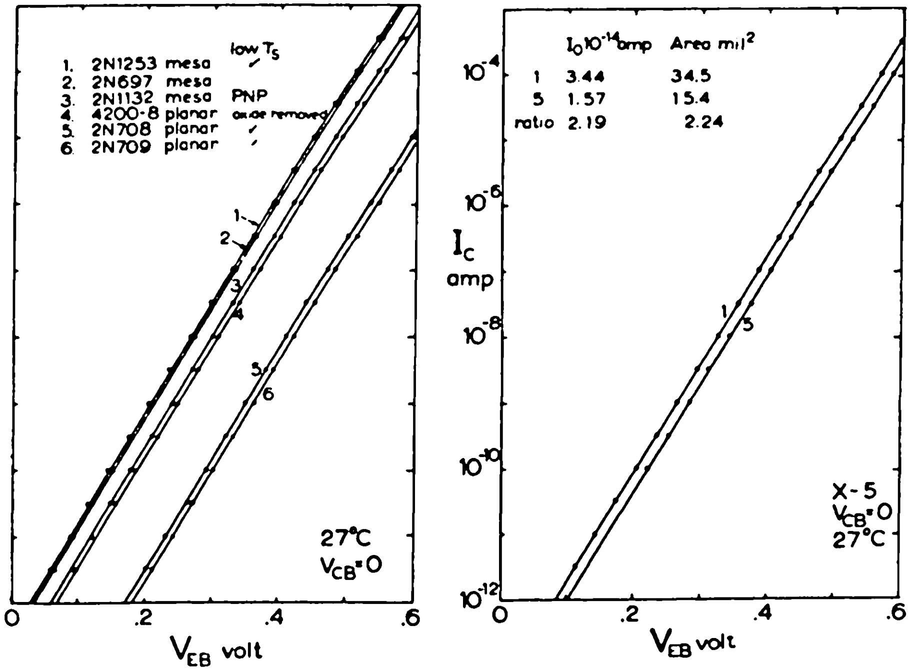
The true exponential dependence of the collector current, I_C, on the emitter-base junction voltage, V_{EB}, is illustrated by the experimental data of many silicon BJT transistors given in Fig.737.1(a). These were taken on state-of-the-art double-diffused Si mesa and planar transistors manufactured in 1962 [737.3]. Figure 737.1(b) gives another proof of universality using two double-diffused n/p/n Si BJTs with different emitter areas but the same collector areas. The straight lines of $\log_e I_C$ vs V_{EB} extend over nearly ten decades of the collector current. This unique universal property has been used to build exponential and logarithmic amplifiers, and energy-gap (bandgap) reference for nearly temperature independent constant voltage sources. The extensive experimental data and interpretations given in the original 1962 Sah paper [737.3] helped to show the importance of the 1970 proposal of using the Gummel number as an experimentally readily accessible BJT design and characterization parameter [737.4].

A consequence from a thin base layer is the very low recombination loss in the base as indicated by the $X_B^2/(2D_B\tau_B)$ term in $\alpha_B = 1-X_B^2/(2D_B\tau_B)$ of (737.2) when $X_B << L_B$. Equation (372.4) shows that the low-level minority-carrier lifetime (holes in p/n/p) in the base is $\tau_B = \tau_p \approx \tau_{p0} = 1/(c_{pe}^{t}N_{TT})$. Thus, the fractional recombination loss in the base is proportional to the density of the recombination centers, N_{TT}, as indicated by

$$1 - \alpha_B = X_B^2/(2D_B\tau_B) = t_B/\tau_B = (X_B^2/2D_B)c_{pe}^{t}N_{TT}. \qquad (737.10)$$

However, j_B of (737.3) indicates that the collector current, (737.9), is independent of base recombination because the diffusion-recombination length, L_B, is cancelled out when the base is very thin. The physics is easy to understand. For low base recombination or thin base, the collector and the emitter currents are essentially determined by the minority carrier diffusion flux from the emitter to the collector since few minority carriers are lost by recombination in the base. Thus, the minority carrier concentration, $P(x)$, drops linearly from the emitter-base junction edge, $x=0_B$, towards the collector-base junction edge, $x=X_B$, as indicated in Figs.733.1(b) and (d). This is given by $P(x) = P(0_B)\cdot[1-(x/X_B)] + P_B$ where $P(0_B) = P_B\cdot[\exp(qV_{EB}/kT) - 1]$.

Base recombination loss will cause the declining straight line to droop or concave downwards as indicated in Fig.733.1(d) because the hole diffusion current (or dP/dx) injected by the emitter into the base at $x=0_B$ is smaller at the collector, $x=X_B$, due to recombination loss in the base. The corresponding electron (majority carrier) current from electron-hole recombination in the base is the base current flowing laterally out to the base terminal. This was the very current that gave rise to the voltage drop along the base resistance which we computed in the preceding section, 736.



(a) (b)

Fig.737.1 Experimental $\log_e I_C$ vs V_{EB} showing linearity over nearly 10 decays of collector currents. (a) Data from engineering and production Si n/p/n and p/n/p mesa and planar BJTs of the 1960 vintage. (b) Two special Si n/p/n Si planar BJTs with different emitter areas and identical collector area. From [737.3].

Gummel Number in the Quasi-Neutral Emitter Layer

Recombination of electrons injected from the n-type base layer into p-type emitter layer with the holes in the p-type emitter layer will reduce the transistor current gain also. There are three recombination sites or layers in the emitter: the emitter-base junction space-charge layer, the emitter quasi-neutral layer, and the metal/emitter or conductor/emitter contact interfacial layer. The metal/emitter interfacial layer is often combined with the quasi-neutral layer. The losses in the three layers are accounted for by the emitter injection efficiency, γ_E, defined in (734.9B). γ_E contains two terms which represent the losses at the three recombination sites. j_{EB} represents the recombination loss in the emitter-base junction space-charge layer. j_E contains the remaining two recombination losses, one at the recombination centers in the quasi-neutral emitter layer and the other, in the metal/emitter contact interfacial layer.

The electron recombination current in the p-type emitter, described by the coefficient j_E and given by (733.21A), can be simplified if the emitter is very thin compared with the emitter diffusion length, $X_E << L_E$. This is sometimes known as the **transparent emitter** in solar cell terminology. It is transparent to the optically generated carriers which can pass through the emitter with little loss to recombination. It is also a condition commonly met in a high gain transistor. This simplification is similar to that made for j_B in the base layer, (737.3), which gave us the Gummel Number of the quasi-neutral base layer. For the emitter layer,

$$\begin{aligned} j_E &= (qD_E N_E/L_E)\,\mathrm{ctnh}(X_E/L_E) \\ &\approx (qD_E N_E/LE)(L_E/X_E) \\ &= qD_E N_E/X_E = qD_E n_i^2/(N_{AA} X_E) \qquad (737.11) \end{aligned}$$

where $N_{AA}X_E$ can be called the **emitter Gummel Number**. Again it can be computed more exactly by an integral if the dopant impurity concentration varies with position in the emitter layer as is generally the case. In addition, the energy gap and the mobility also vary with position in the emitter layer due to the high concentration of dopant impurity atoms incorporated into the emitter layer. Thus, a more general expression taking into account these position dependences is

$$j_E = \frac{q}{\displaystyle\int_{0_E}^{X_E} [N_{AA}(x)/n_i^2(x) D_E(x)]\,dx} \qquad (737.12)$$

where $n_i^2(x)$ takes into account the energy gap shrinkage due to heavy doping or different emitter and base materials such as the heterojunction bipolar transistors to be described in sections 77n.

Emitter Injection Efficiency Calculated from the Gummel Numbers

Let us consider an example using the simpler result of the emitter Gummel number given by (737.11). The ratio j_E/j_B appears in the emitter efficiency given by (734.9B)

$$\gamma_E = j_B/(j_E+j_{EB}+j_B) = 1/[(j_E/j_B) + (j_{EB}/j_B) + 1]. \qquad (737.13)$$

Using both the thin emitter and thin base approximations, (737.3) and (737.11), we have

$$j_E/j_B = (D_E/D_B)(N_{BB}X_B/N_{EE}X_E) = (D_E/D_B)(G_B/G_E) \qquad (737.14)$$

where the concentrations are the majority carrier or dopant impurity concentrations in the p/n/p base and emitter layers respectively, $N_{BB}=N_{DD}$ and $N_{EE}=N_{AA}$. This shows that the j_E/j_B ratio is inversely proportional to the ratio of the Gummel number, G_B/G_E.

For a well-designed Si transistor, $D_E = 1\text{cm}^2/\text{s}$, $D_B = 10\text{cm}^2/\text{s}$, $N_{BB} = 10^{17}\text{cm}^{-3}$, $N_{EE} = 10^{19}\text{cm}^{-3}$, and $X_B = X_E = 10^{-4}\text{cm}$, then

$$j_E/j_B = (1/10)(10^{17}/10^{19}) = 1/1000. \qquad (737.15)$$

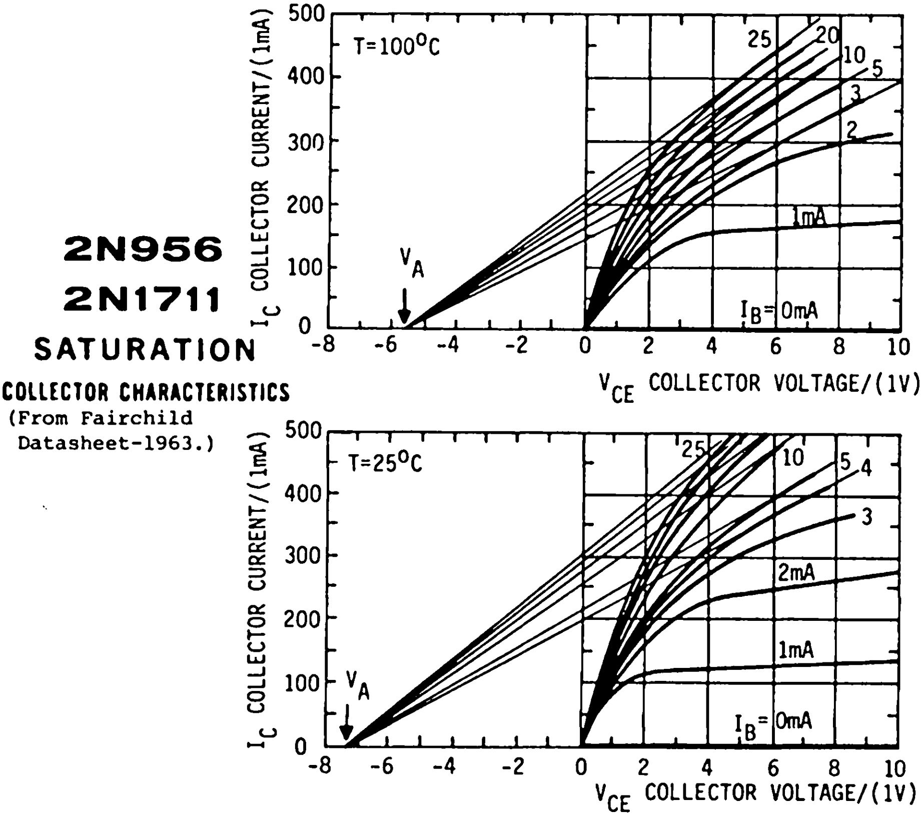
This is a very efficient emitter of minority carriers and it is necessary.

738 Bias Dependences of the D.C. Parameters of BJT

The four 2-port d.c. BJT parameters (I_{ES}, I_{CS}, α_F, and α_R) are not constants as assumed in the original Ebers-Moll transistor equations. The extended Ebers-Moll model or the SNS transistor equations takes into account two of the bias dependences from one neglected physical mechanism: the two currents due to recombination-generation of electrons and holes at traps located in the emitter-base and collector-base space-charge layers. Other bias dependences arise from the modulation of the **electrical thickness** of the layers (quasi-neutral base-layer thickness) and material property (base layer conductivity) by the applied voltage (the **Early effects**) and by the resultant high current (the **Kirk effects**) which alters the Early effect and modulates the base and collector conductivities. In the order of increasing applied voltage and then increasing current density, we shall describe four effects in the following sequence: the Early effect (electrical base thickness modulation by the collector-base junction voltage), the SNS effect (emitter-base space-charge layer recombination current), the base conductivity modulation effect (high injection level in the electrical base), and the Kirk effects (electrical base layer thickening with increasing current or the reverse Early effect). These all occur in a one-dimensional BJT which we shall describe in this section. In real two- and three-dimensional BJTs, some of these effects are modified slightly and others significantly by fringe electric fields and current crowding. In practice they can be accurately modeled by lumped non-interactive 1-d models or distributed models derived from cascading lumped non-interactive 1-d models.

The Early Effects in BJT and Lowly-Doped Collector

The d.c. output characteristics of BJT in both the common-base (I_C-V_{CB}) and common-emitter (I_C-V_{CE}) configurations show a large slope or conductance. The Shockley diode theory predicts zero slope or zero conductance when V_{CB} is at a large reverse value since the Shockley diode current saturates. The reverse current of SNS diode theory does give a finite conductance since the reverse current is proportional to the collector-base junction space-charge layer thickness, $X_{CB} = \surd[2\varepsilon_s(V_{bi}+V_R)/qN_M]$ from (531.1). But the SNS diode conductance is too small to account for the observed output conductance across the collector-base terminals of a BJT. The large experimental output conductance is more visual in the common-emitter configuration than in the common-base configuration since the slope or conductance is increased by approximately β_F. Figure 738.1 shows the expanded I_C-V_{CE} curves of a high-gain Si n/p/n Si BJT with accentuated Early effect. Si p/n/p BJTs can have an even larger Early effect due to the lower-doped n-base necessary to give high hole mobility in the base for high beta.



2N956
2N1711
SATURATION
COLLECTOR CHARACTERISTICS
(From Fairchild Datasheet-1963.)

Fig.738.1 The common-emitter output characteristics of a high-gain Si n/p/n transistor showing a large Early effect or large output conductance in the low-current range of the forward active mode. The high conductance can be characterized by the Early voltage, V_A, indicated in the figure.

The increasing collector current and large output conductance are due to the decrease of the electrical (or quasi-neutral) base thickness with increasing magnitude of the applied reverse collector-base voltage. This was first and successfully modeled by Early in 1952 [738.1] and is known as the Early effect. There are two Early effects. (i) The output or collector current is higher and has a larger dependence on the applied collector-base junction voltage than predicted by the theory just described, even at very low currents. (ii) The lateral base resistance increases owing to decreasing base thickness with increasing reverse bias applied to the collector-base junction. These have been known as the **base thickness modulation effects** or base width modulation effects. The term 'width' has been used historically which correctly described the physical structure of a grown-junction BJT, Fig.710.1(c), in a Ge bar. But the geometry of the Si diffused BJTs used in modern integrated circuits, Figs.710.1(f) and Figs.720.2(a) and (b), dictates the use of the term 'width' for line width of the emitter-base junction stripe, and of the emitter and base interconnect metal stripes. We shall also add the descriptive, electric, and call this the **electric-base-thickness modulation** in order to distinguish it from the **geometric** or **metallurgical base thickness** which is not modulated and normally fixed at a constant value determined by the transistor fabrication conditions. The geometrical or metallurgical base thickness can change under extreme operational stress, such as very high temperature and/or very high electric fields which cause the dopant impurity atoms or some deleterious ions (such as hydrogen, sodium or lithium) to move into the critical layers of the transistor. The BJT would then fail.

Base thickness modulation can be minimized or almost eliminated if we note that a major part of the space-charge layer lies on the lower-doped side of a p/n junction. In a double diffused n++/p+/n transistor, most of the space-charge layer lies on the collector side (or n-side) of the base-collector p+/n junction, so the base layer modulation is small. But if the base is very thin, the small fraction that penetrates into the base may still be a significant percentage of the base thickness. Imbedding a lowly doped layer, not quite intrinsic but commonly called the **intrinsic layer** or **i-layer**, on the p+ or n side of the collector-base p+/n junction would nearly fix the thickness of the base layer and essentially eliminate the Early effects. This **lowly-doped collector** (LDC) was proposed and demonstrated by Early in 1954 [738.2] in Ge and Si BJTs with the **p/n/i/p** or **n/p/i/n** structure.

[738.1] James M. Early, "Effects of space-charge layer widening (thickening) in junction transistors," Proc. IRE, 40(11), 1401-1406, Nov. 1952

[738.2] James M. Early, "The PNIP and NPIN junction transistor triode," Bell Sys. Tech. Jour., 33(5), 517-527, May, 1954.

Most of the applied V_{CB} appears across the i-layer giving high electric field. Thus, the minority carriers drift through the i-layer at the phonon-scattering-limited velocity, 10^7cm/s, is 10^{-12}s if the i-layer is 0.1μm thick. The pinp or npin structure has two additional advantages: an increase of the breakdown voltage and decrease of the capacitance of the collector-base junction. The i-layer requires additional processing steps and it has been incorporated in BJT designs for high voltage applications. Early's 1954 idea of LDC (lowly-doped collector) using an i-layer in the BJT collector-base junction was recently (1980) used to design the drain/channel junction of MOSTs, known as the **lightly doped drain** (LDD). Its objective was exactly the same as that of the BJT: to eliminate the channel length shortening effects.

The base thickness modulation by the collector-base voltage that gives the Early effects can be analyzed using the abrupt collector-base p/n junction model. Its total space-charge layer thickness is given by (531.1)

$$X_{CB} = \sqrt{2\varepsilon_s(V_{CBbi}-V_{CB})/qN_{MCB}} \tag{738.1}$$

where $V_{CB} < 0$ for a reverse bias on the p/n collector-base junction. V_{CBbi} is the built-in potential barrier height of the collector-base n/p junction given by $V_{CBbi} = (kT/q)\log_e(N_{BB}N_{CC}/n_i^2)$ and $N_{MCB} = N_{BB}N_{CC}/(N_{BB}+N_{CC})$ is the geometrical mean of the dopant impurity concentration in the base, N_{BB}, and in the collector, N_{CC}. The fraction of the collector space-charge layer that penetrates into the base layer is given by the ratio $N_{CC}/(N_{BB}+N_{CC})$. Thus, the V_{CB} dependence of the electric base thickness is given by

$$X_B = X_{BO}(\text{Geometrical}) + \Delta X_{BO} \tag{738.2}$$

where

$$\Delta X_{BO} = -\,[N_{CC}/(N_{BB}+N_{CC})]X_{CB} \tag{738.2A}$$

$$= -\,\sqrt{[2\varepsilon_s(V_{CBbi}-V_{CB})/qN_{BB}]N_{CC}/(N_{BB}+N_{CC})} \tag{738.2B}$$

which decreases when the reverse bias magnitude, $|-V_{CB}|$, increases. This verifies the discussion just made: to reduce or eliminate base thickness modulation, N_{CC} is made much smaller than N_{BB}, then $N_{CC}/(N_{BB}+N_{CC}) \approx (N_{CC}/N_{BB}) << 1$ so that the space-charge layer is almost entirely in the collector layer or collector body and little in the base layer.

The Early effect on the base resistance is immediately obvious since the base resistance is inversely proportional to the electrical base thickness. The one-lump base resistance given in section 736 is

$$r_{b'} = (\text{Resistivity})\text{x}(\text{Length}) + (\text{Area})$$

$$= (1/q\mu_B N_B B)(W_B/2) + (X_B \text{x} L_B) \propto 1/X_B. \tag{738.3}$$

Thus, the base resistance increases with increasing reverse bias applied to the collector-base junction. The factor 1/2 assumes two base contact stripes.

To determine the base layer thickening effects on the collector current and the output conductance (or resistance) we note that the electrical base thickness, X_B, is the very base thickness parameter in the BJT equations. Its decrease with increasing $|-V_{CB}|$ would increase the collector current and the output conductance. Its decrease would also increase the alpha and beta due to thinning of the base. These can be illustrated mathematically by isolating out the X_B dependences of the two contributing factors in alpha, α_B and j_B. From $\alpha_F = \alpha_B j_B/(j_B + j_{BE} + j_E)$ of (734.7B) and the approximations of α_B in (737.2), and j_B in (737.3), we have

$$\alpha_F = \alpha_B \gamma_E \approx \frac{1 - X_B^2/(2D_B\tau_B)}{1 + [(j_E + j_{EB})/qD_B P_B] X_B} \tag{738.4}$$

which shows that $\alpha_F \approx 1 - A_1 X_B - A_2 X_B^2$. The linear term usually dominates. The quadratic term is very small compared with 1 in high gain transistor since $1 - A_2 X_B^2$ comes from the base transport factor, $\alpha_B = 1 - A^2 X_B^2$. Thus, as X_B decreases due to increasing reverse bias $|-V_{CB}|$, the alpha will increase towards unity. But, since alpha for a high gain transistor is already close to unity, the rise is small and not visible in the common-base I_C-V_{CB} curve. Similarly, the increase of the slope or the output conductance in the common base configuration is also small and not visible.

However, in the common-emitter configuration, the increases of collector current and output conductance are magnified by the current gain, beta, or $h_{FE} = \beta_F = \alpha_F/(1-\alpha_F)$, which is given by

$$\beta_F = h_{FE} = \frac{1}{(X_B^2/2D_B\tau_B) + [(j_E + j_{EB})/qD_B N_B] X_B} \tag{738.5}$$

$$\approx [qD_B N_B/(j_E + j_{BE})]/X_B. \tag{738.5A}$$

This shows that beta is inversely proportional to the electrical base thickness and hence it is strongly dependent on the electrical base thickness.

The collector current, $I_C = I_{CEO} + \beta_F I_B$, would be also strongly dependent on the collector-emitter voltage, so would be the collector output conductance or the slope of the I_C-V_{CE} curve. Using the above and $X_B(V_{CB})$ given by (738.2),

$$I_C = I_{CEO} + \beta_F I_B \tag{738.6}$$

$$\approx I_{CEO} + \beta_{F0}(I_B/nV_A)(V_A + V_{CBbi} - V_{CB}). \tag{738.6A}$$

V_A is known as the **Early Voltage**. β_{F0} is the forward beta at a low V_{CB} or V_{CE} where the Early effect is unimportant. n is the exponent of the voltage dependence (V_{CE} or V_{CB}) of the collector-base junction space-charge layer thickness, given by $X_{CB} \propto (V_{CBbi} - V_{CB})^n$. The Early Voltage is given by

$$V_A = V_{CB1}(X_{B0}^2/X_{CB1})^n \tag{738.7}$$

$$= X_{B0}^2 N_{BB}(N_{BB} + N_{CC})/(2\varepsilon_s N_{CC}) \qquad (\text{if } n=2) \tag{738.7A}$$

The approximation given by (738.6A) is derived by an expansion near $X_B \to 0$ or when the Early effect is very large. The explicit result given by (738.7A) for the Early Voltage is for an abrupt collector p/n junction with constant dopant impurity concentration on both sides of the junction. The output conductance is obtained by taking the derivative of (738.6A) and is

$$g_o = |\partial I_C/\partial V_{CB}|\Big|_{I_B} = \beta_{FO}|I_B/nV_A| = I_C/nV_A. \qquad (738.8)$$

These results show that the Early effect greatly increases the output current and conductance. The simple model of the Early effect using the Early Voltage, V_A, is consistent with the experimental data shown in Fig.738.1.

The SNS Effects in BJT (Alpha and Beta Fall-off at Low Current)

Shockley's transistor equations predict a bias-independent constant current gain or current transfer ratios, alpha and beta, because all of the Shockley currents vary with the Boltzmann factor, exp(qV/kT). In practice, the current transfer ratios of BJTs drop off at both low and high currents. The fall-off of beta or h_{FE} is particularly severe at low current as indicated in Fig.738.2 (for n/p/n) while the corresponding alpha fall-off is less, because beta fall-off is magnified by large beta. The fundamental cause of this fall-off was shown to be due to the recombination of holes with electrons in the emitter-base junction space-charge layer [737.3]. Recombination prevents holes from getting injected into the n-type base layer from the p-type emitter layer (for p/n/p). Recombination in the emitter-base junction space-charge layer was included in the SNS transistor equations. Hence, the SNS transistor equations will predict the alpha and beta drop-off at low current.

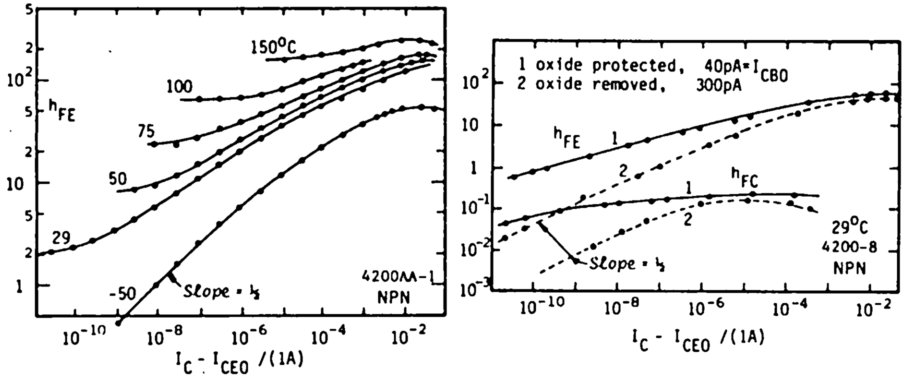


Fig.738.2 The current dependence of the common-emitter d.c. current gain in double diffused silicon bipolar transistors. From [737.3].

To obtain the voltage and current dependences of alpha and beta, we focus on the emitter injection efficiency, γ_B, in the SNS equations. Since the space-charge layer recombination current varies as $\exp(qV_{EB}/2kT)$ while the diffusion current injected into either the base or emitter quasi-neutral layers varies as $\exp(qV_{EB}/kT)$, one would expect a strong dependence of the emitter injection efficiency on the emitter-base junction voltage.

An analytical expression can be obtained from the ratio j_B/j_{BE} to give the emitter-base voltage dependences which can be further worked on to give the collector or emitter current dependencies. The emitter injection efficiency defined by (734.9B) is

$$\gamma_E = j_B/(j_E + j_{EB} + j_B) = [1 + (j_E/j_B) + (j_{EB}/j_B) + 1]^{-1}. \qquad (738.9)$$

Using the defining equation for j_{EB} given by (733.21B), and the approximation for j_B in high gain BJTs given by (737.3), we have

$$\frac{j_{EB}}{j_B} = \frac{(qn_iX_{EB}/\tau_{EB})}{(qD_Bn_i{}^2/N_{BB}X_B)} \cdot \left[\frac{\exp(qV_{EB}/2kT) - 1}{\exp(qV_{EB}/kT) - 1}\right] \qquad (738.10)$$

$$= \left[\frac{X_BX_{EB}N_{BB}}{D_B\tau_{EB}n_i}\right] \cdot \left[\frac{1}{\exp(qV_{EB}/2kT) + 1}\right] \qquad (738.11)$$

$\tau_{EB}=\tau_p+\tau_n$ is the lifetime in the emitter space-charge layer. $N_{BB}=N_{DD}$ is the base impurity dopant concentration. This ratio shows that the emitter-base space-charge-layer recombination rate continues to decrease exponentially with increasing emitter-base voltage V_{EB} as $\exp(-qV_{EB}/2kT)$. This factor comes from the fact that the carrier concentration increases with V_{EB} as $\exp(qV_{EB}/2kT)$ in the space-charge layer and as $\exp(qV_{EB}/kT)$ in the quasi-neutral layer, and that the recombination rate is proportional to the carrier concentration.

Using the numerical values assumed previously and 1μs for the electron and hole lifetimes, then the coefficient in front of the exponential in (738.11) has the numerical value of

$$(X_BX_{EB}/D_B\tau_{EB})\cdot(N_{BB}/n_i) =$$

$$(0.1\text{x}1/10\text{x}10^{-6}\text{x}10^8)\cdot(10^{17}/10^{10}) = 1000. \qquad (738.12)$$

Thus, J_{EB}/j_E drops to 1 when V_{EB} = (2kT/q)log$_e$(1000) = 13.86(kT/q) = 357mV. At this voltage, γ_E=0.5, which is very low and this BJT will not amplify. The corresponding collector current density is 1.6x10^{-5} A/cm^2 and the collector current is 8x10^{-9} A for a 250μm diameter emitter or an area of 4.9x10^{-4} cm^2. However, several specific transistor devices with very important applications depend on the property of increasing emitter injection efficiency with increasing emitter current whose normal operating condition occurs at a low alpha or low gamma. Thus, $\gamma_E=0.5$ may be considered a threshold and the emitter-base junction voltage at this point may be termed the **emitter injection threshold voltage** whose general definition is

$$V_{EBit} = (2kT/q)\log_e[X_E X_{EB} N_{BB}/D_B \tau_{EB} n_i]. \qquad (738.13)$$

One of these applications is the occurrence of a negative resistance in the p/n/p or n/p/n transistor characteristics due to the internal positive feedback loop from the interband impact generation of electron-hole pairs in the collector space-charge layer by the injected minority carriers from the emitter. This will be discussed in section 739. A second application is the bistable property of the p/n/p/n 4-layer diode and transistor switch to be discussed briefly discussed in 739. The latter has been known as the silicon-control-rectifier (SCR) and widely used in power control applications. A third application is the 4-layer p/n/p/n action which causes destructive latch-up of CMOS to the low-voltage/high-current on-state since the p-channel (pnp) and an n-channel (npn) MOS transistors in the CMOS form an extended p/n/p/n structure. The latch-up can be triggered by noise.

In small-signal amplifier and switching applications of BJTs, a more useful definition is the condition at which recombination in the emitter space-charge layer is no longer very important, for example, 0.1% of the emitter recombination. This condition can also be computed. Letting $j_{EB}=j_B/1000$, then

$$V_{EB} = (2kT/q)\cdot\log_e(1000\text{x}1000) = 0.714 \text{ V} \qquad (738.14)$$

and

$$J_C = 1.6 \text{ A/cm}^2 \qquad (738.15)$$

or

$$I_C = 8.0\text{x}10^{-4} \text{ A} = 0.8 \text{ mA} \qquad (738.16)$$

for an emitter-base junction area of $5\text{x}10^{-4}\text{cm}^2$.

The current dependence of the emitter injection efficiency is necessary in applications of the BJT theory to the analysis of the negative resistance characteristics in BJT in the common-emitter configuration and in p/n/p/n devices. The collector current dependences of α_F and h_{FE} are especially useful and needed since in circuit designs using the Ebers-Moll model, the collector current is often specified. To obtain the collector current dependences, we assume a reasonable high current or $V_{EB} > 4kT/q \simeq 100$ mV so that we can drop the term 1 compared with $\exp(qV_{EB}/2kT)$ and $\exp(qV_{EB}/kT)$. Thus, from the second SNS transistor equations, (733.20) or (734.2), we have

$$I_C \simeq -\alpha_B j_B A_E\cdot\exp(qV_{EB}/kT). \qquad (738.17)$$

This can be used to eliminate V_{EB} in (738.11) to give

$$\begin{aligned} j_{EB}/j_B &= (qn_iX_{EB}/\tau_{EB}j_B)/[\exp(qV_{EB}/2kT)+1] \\ &\propto (qn_iX_{EB}/\tau_{EB}j_B)\sqrt{\alpha_BA_E/(-j_BI_C)} && (738.18) \\ &\propto \frac{(qX_{EB}/\tau_{EB})\sqrt{\alpha_B}}{\sqrt{(qD_B/N_{DD}X_B)}\sqrt{(-I_C/A_E)}} && (738.19) \\ &\equiv \sqrt{J_{Cit}/(-J_C)} && (738.20) \end{aligned}$$

where J_{Cit} is the collector current at the injection threshold of the emitter-base junction when the emitter-base junction recombination current is equal to the emitter injection current into the base, $j_{EB}=j_B$, and it is defined by

$$J_{Cit} = (qX_{EB}/\tau_{EB})\sqrt{\alpha_B(qD_B/N_{DD}X_B)}. \qquad (738.21)$$

Note an important result. The strong temperature dependence of space-charge layer recombination from n_i is cancelled out of J_{Cit}. Thus, the dependence of the emitter injection efficiency on the collector current has a much smaller temperature dependence. The beta fall-off is obtained from (735.28), (734.7B), and (734.9B).

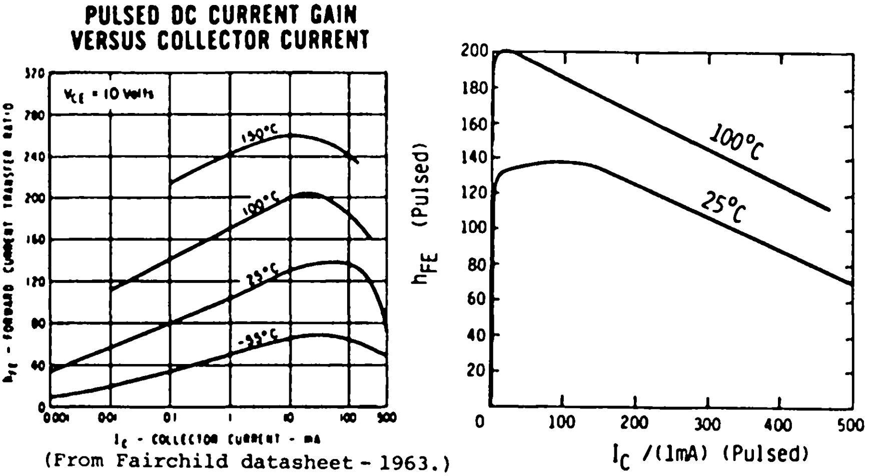
$$\begin{aligned} h_{FE} &\equiv \beta_F = \alpha_F/(1-\alpha_F) = \alpha_B\gamma_E/(1-\alpha_B\gamma_E) \\ &\propto \gamma_E/(1-\gamma_E) = [(j_E/j_B)+(j_{EB}/j_B)]^{-1} \propto (j_{EB}/j_B)^{-1} \\ &= \sqrt{|J_C|/J_{Cit}}. && (738.22) \end{aligned}$$

The square-root collector current dependence of the beta or d.c. h_{FE} has been observed experimentally. This is shown in Figs.738.2(a) at -50°C and in the transistor labeled 2 in Fig.738.2(b).

In summary, the extended-EM or SNS transistor equations, (734.1) and (734.2), predicts the BJT characteristics at low current densities where the original Ebers-Moll equations fail. The success comes from including the electron-hole recombination in the emitter-base junction space-charge layer which dominates at low current densities. This recombination loss reduces the amplification at low current. It significantly influenced the advancement of low-power and high-impedance bipolar linear integrated circuit during the 1960's, such as the operational amplifiers. For example, the high input-impedance bipolar operational amplifier integrated circuit chip was successfully designed by Widler and mass produced by the National Semiconductor Corporation in the mid-1960s. They recognized the basic SNS mechanism that causes the low-current beta fall-off and were able to minimize its magnitude in order to fabricate high-beta Si BJTs for the input stage of the OP-AMP chip, achieving beta's exceeding 10,000 and input resistances greater than 10^9 ohms.

Alpha and Beta Fall-Off at High Current in BJT

In high-power/high-current and high-speed BJTs, the transistor beta decreases with increasing collector current. This is illustrated by figures of common-emitter current gain versus collector current, h_{FE}-I_C, from data sheets of Si BJT. A typical example is given in Fig.738.3. An inverse linear dependence, $h_{FE} \propto 1/I_C$, is commonly observed.



(From Fairchild datasheet - 1963.)

Fig.738.3 Pulsed DC current gain vs collector current of a double-diffused Si planar n/p/n BJT 2N1711 (same as 2N696) first manufactured by Fairchild Semiconductor around 1962.

Two effects account for this beta drop-off at high currents: (1) the injected minority carrier concentration becomes comparable to and then exceeds the dopant impurity concentration, $P(x) > N_{BB}$ where $N_{BB} = N_{DD}$ in p/n/p or p/n/i/p and $N_{BB} = N_{AA}$ in n/p/n or n/p/ν/n BJTs; and (2) current crowds towards the perimeter of the emitter-base junction due to voltage drop along the base resistance.

Effect (1) causes the injected hole density into the base to rise slower with the emitter-base voltage, as $\exp(qV_{EB}/2kT)$, while the injected electron density into the emitter is still increasing at the low level condition, $\exp(qV_{EB}/kT)$. Thus, an increasingly larger fraction of the emitter current is carried by minority carrier injection into the emitter from the base or j_E/j_B is increasingly larger. Its on-set condition can be controlled by the designed-in material variations such as the lowly-doped i-region and ν-region in high-voltage BJTs. Still, this is a one-dimensional effect, however, its onset will be modified by two- and three-dimensional geometry. This effect is known as the **base high-injection-level** or **base high-level** effect. It will be analyzed shortly since it can be treated by simple one-dimensional arithmetic.

Effect (2) is a two-dimensional effect. The voltage drop along the base resistance lowers the junction voltage towards the middle of the emitter-base junction below the terminal V_{EB}. The higher forward voltage at the perimeter of the emitter-base junction gives higher current density, resulting in current concentration or crowding at the perimeter of the emitter-base junction. The higher current density would set the perimeter region into the high injection level condition described by effect (1) sooner than the middle region of the emitter-base junction. The fringing current at the perimeter may also reduce the gain due to a larger effective base thickness or longer minority carrier path to the collector junction but this effect is secondary compared with the high injection level effect from $\exp(qV_{EB}/2kT)$.

The onset (actually a smooth and gradual beginning) of the high injection level condition in the base layer can be derived quantitatively. The condition at which the low level assumption begins to break down is when the minority carrier concentration in the base layer is about 10% of the majority carrier concentration. Consider a base doping of 10^{17} phosphorus-donor/cm^3. Since the minority carrier concentration (hole) is the highest at the base-side edge of the emitter space-charge layer, which is given by the boundary condition

$$P(0) = P_N \exp(qV_{EB}/kT), \tag{738.23}$$

then the condition at which the low level approximation begins to fail is

$$P(0) = P_N\exp(qV_{EB}/kT) = N_N/10 = N_{BB}/10 \tag{738.24}$$

$$= 10^{17}/10 = 10^{16}\ \text{cm}^{-3}. \tag{738.24A}$$

The corresponding emitter-base junction voltage, to be known as the **high-level threshold voltage**, is then

$$V_{EBhl} = (kT/q)\cdot \text{loge}(N_N/10P_N) = (kT/q)\cdot \log_e(N_{DD}{}^2/10n_i{}^2) \tag{738.25}$$

$$= 0.02585\cdot \log_e(10^{17\text{x}2}/10\text{x}10^{20}) = 0.7738\ \text{Volt}. \tag{738.25A}$$

The corresponding emitter or collector **high-level threshold current density** is

$$J_{Ehl} \simeq J_{Chl} = (qD_BP_B/\tau_B)\text{ctnh}(X_B/L_B)[\exp(qV_{EB}/kT) - 1] \tag{738.26}$$

$$\simeq (qD_B/X_B)[P_B\cdot \exp(qV_{EB}/kT)] = (qD_B/X_B)[N_{DD}/10] \tag{738.26A}$$

$$= 1.6\text{x}10^{-19}\text{x}10\text{x}10^{17}/10\text{x}10^{-4} = 160\ \text{A/cm}^2. \tag{738.26B}$$

For a micron-size silicon BJT in a VLSI chip, its emitter area may be 2μmx5μm or smaller, which would give a collector current at the onset of high level of 16μA. For a larger silicon transistor of 10-mil or 250μm diameter, such as the first double-diffused high-speed Si planar n/p/n 2N706, the area is $5\text{x}10^{-4}\text{cm}^2$ and the current is about 80 mA. The first production Si planar n/p/n transistor 2N696 had a 15-mil emitter diameter and a current of about 180mA (agrees with Fig.738.3), while its nominal rating is about 100mA.

To predict the alpha and beta fall-off at high currents, the low level boundary condition for the concentration of the minority carriers entering the quasi-neutral base layer, (738.23), must be modified. At low levels, defined as when the injected minority carrier concentration is much less than the dopant impurity concentration, the majority carrier concentration is approximately equal to or only minutely greater than the dopant impurity concentration. When the injected minority carrier concentration level increases and exceeds the dopant impurity concentration, the majority carrier concentration also increases in order to maintain electrical neutrality or quasi-neutrality. In general, the product of the electron and hole concentrations is still given by

$$P(x)N(x) = n_i^2\exp[qV(x)/kT] \qquad (738.27)$$

and V(x) is the difference between the quasi-Fermi potential of electron and hole described in section 331 or just the voltage difference. Together with the quasi-neutral condition

$$\rho(x) = q[P(x) - N(x) - N_{AA} + N_{DD}] \approx 0 \qquad (738.28)$$

the electron and hole concentrations can be obtained. This problem was already solved in section 242 for lowly doped or nearly intrinsic semiconductor at equilibrium where $V(x)=0$ and $pn = n_i^2\exp[qV(0)/kT] = n_i^2$. Replacing $pn=n_i^2$ in the equilibrium solution by $n_i^2\exp(qV/kT)$ of (738.27), then the hole concentration in n-type silicon is

$$P(x) = (1/2)[- N_{DD} + \sqrt{N_{DD}^2 + 4n_i^2\exp[qV(x)/kT]} \qquad (738.29)$$

$$\approx (n_i^2/N_{DD})\exp[qV(x)/kT] \qquad (\text{if } P(x) < N_{DD}/10) \qquad (738.29A)$$

$$\approx n_i\exp[qV(x)/2kT] \qquad (\text{if } P(x) > 10N_{DD}). \qquad (738.29B)$$

Equation (738.29) reduces to the correct low level boundary condition (738.29A), $P(x)=P_B\exp(qV/kT)$, where $P_B=n_i^2/N_{DD}^2$ given by (738.29A). The high-level boundary condition (738.29B) shows that the rate of rise of the logarithmic hole concentration with voltage, $(d/dV)\log_e P(x)=q/2kT$, is one half of that at low level which is kT/q.

The high injection level condition begins first at the emitter-base-junction edge, $x=0_B$ in Fig.733.1(c), of the quasi-neutral base layer. It penetrates into and eventually engulfs the entire thickness of the base layer as the emitter-base junction voltage increases further. Due to low recombination in the base, its onset and complete penetration occur almost simultaneously (within a few kT/q of increase of V_{EB}) so that the general boundary condition given by (738.29) applies essentially to the entire quasi-neutral base layer. At the far edge of the base layer, $x=x_B$ in Fig.733.1(c), where the collector-base junction space-charge layer begins, the high electric field will sweep out the injected carriers at the phonon-scattering-limited carrier saturation velocity, θ_{SAT}, which is about equal to the thermal velocity, 10^7 cm/s. This is one of the hot carrier effects described in chapter 3. Velocity saturation would limit the carrier concentration at this edge to the value given by $J_C=q\theta_{SAT}P$ or $P=J_C/q\theta_{SAT}$. Under normal current density, this is less than one-tenth of the injected value $P(0_B)$ and can be assumed zero. At high injection levels, this may not be the case; however, its effect is small until the effective diffusion velocity in the quasi-neutral base, D_B/X_B, becomes comparable to θ_{SAT}. Otherwise, diffusion is limiting. This is easy to see since the injected carriers diffuse through the quasi-neutral base first and then enter into the drift layer or the collector-space-charge layer. Since the diffusivity in the base is about 5 cm^2/s, to reach 10^7 cm/s diffusion velocity, the base must be $X_B < D_B/\theta_{SAT} = 5/10^7 =$ 50A, which is not attainable with the present technology. Thus, for all practical purposes, the carrier density at the edge of the collector-base space-charge layer can be considered zero. Exceptions are when the collector is not properly designed in a p/n/i/p or n/p/i/n structure or the BJT is operated outside its intended range, resulting in incomplete penetration or incomplete depletion of the entire lowly-doped collector (LDC) layer. This would cause a large collector series resistance from the unpenetrated or undepleted part of the i-layer. This could happen in the saturation range when the collector-base junction is forward-biased or slightly reverse biased. For a properly designed i-layer, conductivity modulation at high injection level would reduce or eliminate the high collector series resistance.

Three other parameters vary with the injection level. All give **increasing alpha** and **beta** with increasing emitter and collector current densities, exactly opposite to the high current experimental results. We now explain the basic physics of these three effects. (i) The recombination lifetime increases from τ_{p0} to $\tau_{p0}+\tau_{n0}$ as injection or emitter current increases because the delay time of capturing a majority carrier (electron) by a trap as injection level increases becomes comparable to that of the minority carrier (hole) owing to the comparable minority and majority carrier concentrations at high level. This is known as the **ambipolar** or **high-level lifetime**. (ii) The diffusion of the holes across the quasi-neutral n-type base layer is sped up because the aiding electric field created by the gradient of the electrons in the quasi-neutral base at high injection levels which increases the diffusivity from D_p to the **ambipolar diffusivity**, $2D_pD_n/(D_p+D_n)$ [738.3]-[738.4]. It approaches $2D_p$ if $D_n >> D_p$ but remain D_p if $D_n=D_p$. (iii) Under the most ideal assumption of zero majority carrier current or zero recombination, frequently made by academicians but never occurs in practice, the electric field from longitudinal variation of majority carriers (electrons) creates a large drift field and minority carrier (hole) drift current which adds to its diffusion current. At the high injection level limit, the drift current equals the diffusion current. This doubles the total minority carrier current to $J_P=-2D_p dP/dx$. However, this cannot produce a variation of α or β with current because zero recombination means $\alpha=1$ and $\beta=\infty$.

The increasing minority carrier diffusivity with increasing current was first analyzed by Webster of RCA in 1954 [738.5] in a failed attempt to explain the beta rise with current at low current levels and not the beta fall at high currents. Webster's contributions have often been misinterpreted by recent textbook author such as [738.3]. An alternative derivation, given by Gummel and used by Gummel and Poon to analyze high-level effects [738.6], gave the correct voltage dependence of the hole concentration. They used a clever algebraic manipulation under the assumption of zero recombination. This puts severe restriction on the applicability of the result which has been overlooked by later practioners.

[738.3] S.M.Sze, **Physics of Semiconductor Devices**, 2nd Ed., pp.86-87, John Wiley & Sons, New York, 1981.

[738.4] David J. Roulston, **Bipolar Semiconductor Devices**, p.72 and p.227, McGraw-Hill Publishing Co. New York, 1990.

[738.5] W. M. Webster, "On the variation of junction-transistor current amplification factor with emitter current," Proc.IRE 42(6),914-920, June 1954.

[738.6] H.K.Gummel and H.C.Poon, "an integral charge control model of bipolar transistor," Bell System Tech. J., 49(5),827-852, May-June 1970.

It is evident from the foregoing discussion that the principal high-level effect on alpha and beta comes from the boundary condition of electron and hole densities at the emitter-base junction, not the enhanced ambipolar diffusivity (the Webster effect) and lifetime (SRH) in the base layer as stated by some textbook authors. The boundary condition, (738.29), shows a continuous slowdown of $P(x=0)$ or decreasing $dP(x=0)/dV_{EB}$ with increasing emitter-base voltage V_{EB}; or $P(0)$ and $dP(x=0)/dJ_C$ with increasing collector current. It is this slowdown that accounts for the continuous **beta fall-off** with increasing collector current. Increasing diffusivity and lifetime with injection level (Webster and SRH effects) would increase and not decrease the collector current and beta. These would give only a sublinear **beta rise** with J_C at large J_C which saturates to roughly four times the low-level beta. Thus, the rise of Webster diffusivity and SRH lifetime with emitter current are insufficient to account for the experimental beta rise with current at low current. They cannot account at all for the very large thermally-activated increase of the experimental low-current beta with increasing temperature, $\beta_F \propto \exp(E_G/2kT)$.

The four factors just discussed can be taken into account to derive an expression for alpha and beta at high levels and high collector-base bias voltage in the active mode. We shall use the same low-level symbols for the variable high-level lifetime and diffusivity in the base since they are not the key factors while the carrier density boundary condition is. Then, the high level j_B is

$$j_B = q(D_B/X_B)n_i\exp(-qV_{EB}/2kT) \qquad (738.30)$$

which is obtained from using $P(x=0) = n_i\exp(qV_{EB}/2kT)$ instead of $P_B\exp(qV_{EB}/kT)$. The collector and emitter current densities are respectively

$$J_C \approx -\alpha_B j_B[\exp(qV_{EB}/kT)-1] \approx -\alpha_B q(D_B/X_B)n_i\exp(qV_{EB}/2kT) \qquad (738.31)$$

and

$$J_E = q(D_E/X_E)(n_i^2/N_E)\exp(qV_{EB}/kT) + q(D_B/X_B)n_i\exp(qV_{EB}/2kT). \qquad (738.32)$$

Using $J_B=-(J_C+J_E)$, then the CE forward current transfer ratio, h_{FE} or β, is

$$\beta_F = h_{FE} = -J_C/J_B = 1/[1 - (1/\alpha_B) + (D_E X_B^2/qN_{EE}X_E D_B^2)(-I_C/A_E)] \qquad (738.33)$$

$$= 1/[(X_B^2/2D_B\tau_B) + (D_E X_B^2/qN_{EE}X_E D_B^2)(-I_C/A_E)] \qquad (738.33A)$$

$$= 1/[(t_B/\tau_B) + (I_C/I_{CHL})] \qquad (738.33B)$$

where

$$J_{CHL} = I_{CHL}/A_E = qN_{EE}X_E D_B^2/D_E X_B^2 \qquad (738.34)$$

is the collector current density when h_{FE} drops to 1. The emitter-base voltage at this current density is obtained by using this current density in the emitter current expression (738.31). This substitution gives

$$V_{EB-HL} = (2kT/q)\log_e[(D_B/D_E)(X_E/X_B)(N_{EE}/n_i)]. \qquad (738.34A)$$

The key result of (738.33B) is that h_{FE} is inversely proportional to the collector current density at high currents. The data from the 2N1711 data sheet given in Fig.738.3 supports this theoretical current dependency. Equation (738.33A) also shows explicitly the beta rise with increasing base lifetime (SRH effect) and diffusivity (Webster effect). Its geometric dependence is $h_{FE} \propto 1/X_B^2$ (high level) and $1/X_B$ (low level), giving a larger Early effect as injection level increases.

As a numerical example, take the numbers used previously, $X_E=0.1\mu m$, $X_B=1.0\mu m$, $D_B=5cm^2/s$, $D_E=4cm^2/s$, and $N_{EE}=10^{19}cm^{-3}$, then

$$J_{CHL} = 1.6\times10^{-19}\times10^{19}\times10^{-5}\times5^2/(4\times10^{-4}\times2) = 10^4\ A/cm^2$$

and

$$V_{EB-HL} = (2kT/q)\log_e[(5/4)(0.1/1.0)(10^{19}/10^{10})]$$
$$= 37.288(kT/q) = 964\ mV.$$

For a state-of-the-art sub-0.5 micron technology with $X_E=X_B=0.1\mu m$, then

$$J_{CL} = 10^6\ A/cm^2$$

and

$$V_{EB-HL} = 41.89(kT/q) = 1.0829\ V.$$

These are rather high values. Another criterion can also be computed from these results: the carrier concentration for velocity limited current flow at this current density. This is $P \leq P_{HL} \equiv J_{CHL}/q\theta_{MAX} = 10^6/(1.6\times10^{-19}\times10^7) = 6.25\times10^{17}\text{cm}^{-3}$ for velocity saturation to be important. Thus, velocity saturation barely exists in a base of $N_{BB}=6\times10^{16}\text{cm}^{-3}$ when the high level condition sets in at $J_C=10^5$ A/cm^2 in a BJT of $h_{FE}=10$.

In addition to the series collector resistance in a p/n/i/p or n/p/i/n structure improperly designed or operated outside of the designed range, the high injection level effect also increases the majority carrier conductivity in the base which decreases the base resistance. This causes the base resistance to be a function of emitter-base voltage or collector current. The base resistance of a two-stripe base contact was given by (736.6). Replacing N_B by $N=P=n_i\exp(qV_{EB}/2kT)$ and using (738.31) to give $P = J_CX_B/(\alpha_B qD_B) \approx J_CX_B/qD_B$, the base resistance is

$$
\begin{aligned}
r_{b'} &= (1/q\mu_n N)(W_B/4ZX_B) = (W_B/4Z)(kT/qJ_CX_B^2)\alpha_B(\mu_B/\mu_p) \\
&\approx (W_B/4Z)(kT/qJ_CX_B^2) \qquad (738.35)
\end{aligned}
$$

which shows $r_{b'} \propto 1/X_B^2$ at high injection level while $r_{b'} \propto 1/X_B$ at low level.

The Kirk Effects in BJT

When the minority or collector current density is very high and limited by the drift current across the collector-base junction space-charge layer, the charge density of the minority carriers in the space-charge layer may become comparable to the dopant impurity ion concentration on either or both sides of the collector-base junction. For a p/n/p BJT, the effective charge densities in the collector-base junction space-charge layer on the n-base side is then

$$
\begin{aligned}
+\rho_N(x) &= +\,q(P-N+N_{DD}) = +\,q(P-N_{DD}) \\
&= +\,q[\,|J_C/q\theta_{MAX}| + N_{DD}] \qquad > \; qN_{DD}. \qquad (738.36A)
\end{aligned}
$$

On the p-collector side, it is

$$
\begin{aligned}
-\rho_P(x) &= -\,q(P-N-N_{AA}) = -\,q(P-N_{AA}) \\
&= +\,q[N_{AA} - |J_C/q\theta_{MAX}|\,] \qquad < \; qN_{AA}. \qquad (738.36B)
\end{aligned}
$$

Here we have used the magnitude so as not to be confused with the sign of J_C which is negative for p/n/p and positive for n/p/n in the current-reference convention we have employed.

Thus, the effective charge on the base side is increased as shown by (738.36A), which decreases the space-charge layer thickness, increases the electrical base thickness, and lowers the beta. On the collector side, the effective charge is decreased as shown by (738.36B) which would increase the penetration of the space-charge layer into the collector body and reduce any series collector resistance that might not have been designed out. It also increases the space-charge layer thickness on the base side due to a thicker overall space-charge layer which decreases the electric base thickness. The net effect on the base thickness is still that of increasing the base thickness and lowering the beta. This is known as the **d.c. Kirk effect** while the original work by Kirk in 1962 dealt with the high-frequency Kirk effect [738.7]. It is also known as the **electric base push-out effect.** The term 'electric' is added since there is also a geometrical collector-base junction push-out effect due to enhanced impurity diffusion owing to donor-acceptor-vacancy interaction at the high impurity diffusion temperatures during transistor fabrication. Note that at extremely high current densities, (738.36B) shows that the edge of the quasi-neutral base can be pushed out all the way to buried n+ sub-collector in a n+/p/n-/n+sub-collector/Si-substrate BJT structure.

[738.7] C. T. Kirk, "A theory of transistor cut-off frequency, f_T, falloff at high current density," IEEE Trans. Electron Devices, ED-9(3), 164-174, March 1964.

Analytical solution of the Kirk effect can be obtained for a collector-base n/p junction with abrupt interface and constant N_{DD} and N_{AA} using the two effective charge densities given by (738.36A) and (738.36B) to replace qN_{BB} and qN_{CC} respectively in (738.2B). This gives

$$X_B = + X_{B0} - \sqrt{[2\varepsilon_s(V_{CBbi}-V_{CB})/q(N_{DD}+N_{AA})]} \cdot\sqrt{[N_{AA}-|J_C/q\theta_{MAX}|]/[N_{DD}+|J_C/q\theta_{MAX}|]} \qquad (738.37A)$$

which may be written in the following more compact form in order to sort out the influence of the Kirk effect on the Early effect:

$$X_B = X_{B0} + \Delta X_B = X_{B0} + K_K(J_C)\Delta X_{B0} \qquad (738.37B)$$

$$= X_{B0} - K_K(J_C)|\Delta X_{B0}| \qquad (738.37C)$$

where X_{B0} is the geometric base thickness, and ΔX_{B0} is the low-current or zero-current Early base-modulation thickness given by (738.2A) and (738.2B) with $N_{BB}=N_{DD}$ and $N_{CC}=N_{AA}$ for the p/n/p BJT

$$\Delta X_{B0} = - \sqrt{[2\varepsilon_s(V_{CBbi}-V_{CB}/q][N_{AA}/N_{DD}(N_{DD}+N_{AA})]}. \qquad (738.38)$$

The negative sign in (738.38) above means thinning of the base layer. The factor $K_K(J_C)$ is the d.c. Kirk coefficient defined by

$$K_K(J_C) = \sqrt{[1 - |J_C/(q\Theta_{MAX}N_{AA})|]/[1 + |J_C/(q\Theta_{MAX}N_{DD})|]} \quad (738.39)$$

$$\propto [1 - |J_C/q\Theta_{MAX}|(N_{AA}+N_{DD})/2N_{AA}N_{DD}] \quad (\text{Small } J_C) \quad (738.39A)$$

$$\equiv [1 - J_C/J_{KK}] \quad (738.39B)$$

where J_{KK} is the current density for the Kirk effect to reach its highest influence and is defined by

$$J_{KK} = q\Theta_{MAX}(2N_{AA}N_{DD})/(N_{AA}+N_{DD}). \quad (738.40)$$

Normally, the Kirk effect ceases when the i-layer is completely depleted. For a BJT without an i-layer, base punch through occurs when $\Delta X_{B0} = -X_{B0}$.

This result confirms our earlier physical reasoning given before we started to derive the analytical solution given by (738.38), namely, the Kirk effect reduces the Early effect.

An analytical formula of $\beta_F(J_C)$ of the Kirk effect can be obtained using (738.37)-(738.40) in the alpha and beta formulae given by (738.4) and (738.5) for p/n/i/p or n/p/i/n structure when the V_{CB} is insufficiently high to deplete the entire i-layer. Consider the emitter current density range which is sufficiently high to put the i-layer into high-injection level but not high enough to put the n-base or p-base into high-injection level. The emitter-base junction boundary condition is still at low level, $P = P_B\exp(qV_{EB}/kT)$, instead of the high injection level $P = n_i\exp(qV_{EB}/2kT)$. In this case, the Kirk effect decreases the Early V_{CB} effect or reduces the high output conductance by a constant factor at the starting current density when the Kirk effect begins. As the current density increases to a sufficiently high value to completely deplete the i-layer but still below the onset current density of the high-level boundary condition, the Kirk effect will remove the Early effect completely.

At high injection levels, the Kirk effects on the Early effect and on the beta drop-off due to the high level emitter-base boundary condition are stronger since $\beta_F \propto X_B^{-2}$. The analytical result is easily obtained using X_B given by (738.37A) in (738.33A).

The Early and Kirk effects are very important on the high-frequency and high-speed performance of BJT's. Their importance is further marked if the high-frequency or high-speed BJT is to be operated at high voltages and currents and at high power levels. The development of the theories by their discoverers were motivated by the high-frequency and high-speed characteristics more than the d.c. characteristics we have just discussed.

739 Collector Multiplication and Negative Resistance

The negative resistance characteristic in the common-emitter configuration at $I_B=0$ shown in Fig.739.1 (p.790) was the first **multi-mechanism** effect discovered in a semiconductor device that produces a bistable current-voltage characteristics. The fundamental regenerative process responsible is the combination of two mechanisms: (i) the increasing alpha with increasing emitter and collector current due to decreasing recombination loss in the emitter-base junction space-charge layer, and (ii) the multiplication of the injected minority carrier concentration and current in the high electric field of the collector-base junction space-charge layer due to impact generation of electron-hole pairs by energetic electrons and holes. The positive feedback loop is obvious: increasing alpha as the emitter current increases, and the increased injected or emitter current produces still higher collector current by multiplication in the collector-base space-charge layer. The loop is closed when the emitter and collector currents are equal, which is effected by floating the base ($I_B=0$) in the common-emitter configuration. The system becomes unstable and gives a negative resistance when the emitter current, I_E, exceeds the sum of the multiplied emitter current, $M\alpha_F I_E$, and the collector leakage current, I_{CBO}. The excess emitter current drives the loop to instability.

Since the multiplication becomes infinite at the collector-base junction breakdown voltage, BV_{CBO}, the maximum attainable V_{CB} must be lower. As the applied terminal voltage, $V_{CE}(I_B=0)$, increases, the loop gain increases towards unity due to the positive feedback. When V_{CB} ($=V_{CE}-V_{EB}$) reaches this lower voltage, the loop gain is unity, the system becomes unstable, and the current diverges. Thus, the BJT must seek a stable state at a higher current. A stable higher current state can be attained and sustained if the loop gain becomes less than unity at some higher current. This can occur only if (i) the collector-base junction voltage is lowered to decrease the current multiplication and (ii) the current gain has a maximum and decreases at still higher currents (alpha fall-off).

There were attempts in the mid-1950's to use the negative resistance to generate microwave signal by Hamilton, Gibbons and Shockley at Stanford; Statz and associates at Raytheon Research Laboratories; and Webster, Kidd and Hasenberg at RCA Laboratories. However, its main contribution to transistor history has been the deleterious lowering and degradation of the operating voltage of the BJT in the common-emitter circuits. But it also led to the most useful 4-layer silicon controlled rectifier (SCR).

The history of this positive feedback loop and negative resistance traces back to the original theory proposed and rejected by Shockley in 1950 [739.1] to use the collector multiplication in a four-layer, three-junction, p/n/p/n configuration known as the hook collector to explain the greater-than-unity current transfer ratio, I_C/I_B, or alpha observed by Bardeen and Brattain in 1948 when they discovered the transistor effect in the point-contact transistor. This positive feedback probably existed in the space-charge region of the collector point contact without the additional p/n junction as the theory to be worked out in the following paragraph indicates. It was J. J. Ebers who in 1952 [739.2] used a two-transistor circuit model to show the existence of negative resistance and bistability. Ebers' circuit was a complementary pair (a p/n/p and an n/p/p transistor) which he connected to simulate Shockley's four-layer p/n/p/n 'hook collector'.

[739.1] W. Shockley, **Electrons and Holes in Semiconductors**, D. Van Nostrand, Princeton, NJ. 1950, p.110-114. See also Phys. Rev. 78, 294-295 (1950).
[739.2] J.J. Ebers, "Four-terminal p-n-p-n transistors," Proc.IRE,40, 1361-1371, Nov.1952.

This negative resistance characteristic can be easily derived by including the interband impact generation of electron-hole pairs by the minority carriers injected into the space-charge layer of the reverse-biased collector-base n/p junction. These injected minority carriers are accelerated by the high electric field in the collector-base junction space-charge layer to high kinetic energies that exceed the impact generation threshold energy (described in section 536). To take this effect into account, the injected collector current given by $J_P(X_B,V_{EB})$ of (733.9B) is multiplied by the impact multiplication factor M_p which increases with increasing collector-base voltage V_{CB}. At a large reverse bias and high electric field, the collector leakage or dark current is also multiplied and in general by a different multiplication factor, M_c. The modified collector current equation in the common base connection of the Ebers-Moll model, (734.4) or (735.8), now reads

$$I_C = -\alpha_F I_{ES}[\exp(qV_{EB}/kT)-1]\cdot M_p + I_{CS}[\exp(qV_{CB}/kT)-1]\cdot M_c \qquad (739.1A)$$

$$\simeq -M_p\alpha_F I_E + M_c I_{CBO} \qquad \text{if } V_{CB} < -2kT/q. \qquad (739.1B)$$

We shall consider only the floating base situation (curve $I_B=0$ in Fig.739.1), $I_B=I_E+I_C=0$ or $I_E=-I_C$. The finite base current case is described with the four-layer three-junction bipolar diodes and silicon controlled rectifiers or switches. Thus, (734.3) becomes

$$-I_C = I_E \qquad (739.2)$$

$$= I_{ES}[\exp(qV_{EB}/kT)-1] - \alpha_R I_{CS}[\exp(qV_{CB}/kT)-1] \qquad (739.2A)$$

$$\simeq I_{ES}[\exp(qV_{EB}/kT)-1]. \qquad (739.2B)$$

Substituting this into (739.1A) to eliminate V_{EB} and let $V_{CB} << -2kT/q$, then

$$I_C \simeq \alpha_F M_p I_C + I_{CS}[\exp(qV_{CB}/kT)-1]\cdot M_c$$

$$\simeq \alpha_F M_p I_C - I_{CBO} M_c$$

which can also be gotten directly from (739.1B) using (739.2). Thus,

$$I_C = \frac{-M_c I_{CBO}}{1 - \alpha_F M_p} \quad . \tag{739.3}$$

Since $\alpha_F = \alpha_F(I_C)$ and increases with increasing collector current due to diminishing recombination in the emitter space-charge layer while $M_p = M_p(V_{CB})$ and increases with the collector-base junction voltage, the denominator, $1\text{-}\alpha_F M_p$, will decrease with increasing applied voltage. As $\alpha_F M_p \rightarrow 1$, the current given by (739.3) will be mainly determined by

$$\alpha_F M_p = 1 \tag{739.4}$$

since $|I_C| >> |M_C I_{CBO}|$. This is the negative resistance range which is clearly indicated by the $I_B = 0$ curve of the I_C-V_{CE} family given in Fig.739.1.

The entire I-V curve can be computed from (739.3). To demonstrate this and to obtain the analytical solutions of the peak and valley voltage and current values, such as those shown in the $I_B = 0$ curve in Fig.739.1, we shall assume $M_p = M_c$. M_p is the collector multiplication factor of the holes injected by the emitter-base junction, while M_c is the multiplication factor of the thermally generated electrons and holes in the collector space-charge layer. In order to obtain an analytical solution, we also assume that their dependencies on the applied and breakdown voltages are representable by a power law:

$$M_p = M_c = 1/[1 - (|V_{CB}|/BV_{CBO})^n] \tag{739.5}$$

where BV_{CBO} is the collector-base junction breakdown voltage at $I_E = 0$ and n is the numerical exponent of impact multiplication factor which is of the order of 5. We use the low-current approximation, (738.22), due to recombination in the emitter-base junction space charge layer, and the high-current approximation, (738.33B), due to high-level boundary condition. Then, from α_F defined in (734.7B),

$$\begin{aligned} \alpha_F^{-1} &= [1 + (J_E/J_B) + \sqrt{(I_{Cit}/I_C)} + (I_C/I_{CHL})]/\alpha_B \\ &= [1 + (D_E/D_B)G_B/G_E) + \sqrt{(I_{Cit}/I_C)} + (I_C/I_{CHL})]/\alpha_B \end{aligned} \tag{739.6}$$

where $(G_B/G_E) = N_{BB}X_B/N_{EE}X_E$ is the ratio of the Gummel number of the base and emitter quasi-neutral layers given in (737.14); I_{Cit} is the injection threshold current due to electron-hole recombination loss in the emitter-base junction space-charge layer defined by (738.21); and I_{CHL} is the collector current at $\beta_F = 1$ defined by (738.34) due to high-level boundary condition at the emitter-base junction and in the quasi-neutral base layer.

Putting (739.5) and (739.6) into (739.3), the bistable p/n/p diode equation

$$I_C[1 - \alpha_F(I_C)M_p(V_{CB})] = M_c(V_{CB})I_{CS}(V_{CB})$$

is obtained which can be solved to give

$$(|V_{CB}|/BV_{CBO})^n = 1 - \alpha_F - (I_{CBO}/I_C)$$

$$= 1 - (I_{CBO}/I_C) - \alpha_B/[1+(D_EG_B/D_BG_E)+\sqrt{(I_{Cit}/I_C)}+(I_C/I_{CHL})]. \qquad (739.7)$$

This is an explicit equation of the voltage V_{CB} as a function of I_C or an implicit equation of the current I_C as a function of the voltage V_{CB}. An I_C-V_{CB} curve can be readily computed using a handheld calculator by assuming an I_C and computing the V_{CB}. Since the emitter is forward biased, V_{EB} is small compared with V_{CB}, the total terminal voltage is essentially given by V_{CB}.

Three ranges of the I_C-V_{CB} characteristics can be delineated with the help of the experimental data shown in Fig.739.1. These are: (1) the blocking or switch-off range at low currents, (2) the negative resistance range in the medium current range, and (3) the low positive differential resistance region in the high current range. These are all predicted by (739.7). Asymptotic solutions can be obtained since the three characteristic currents, I_{CBO}, I_{Cit} and I_{Chl}, are different from each other by several orders of magnitude. Separating the three ranges by the two currents, I_{PEAK} and I_{VALLEY}, (See Fig.739.1 on p.790) the approximate solutions from the exact voltage-current equation, (739.7), are summarized as follows.

$$(|V_{CB}|/BV_{CBO})^n$$

$$= 1 - (I_{CBO}/I_C) - \alpha_B/[1+ (D_EG_B/D_BG_E) +\sqrt{(I_{Cit}/I_C)} + (I_C/I_{CHL})] \qquad (739.7)$$

$$\approx 1 - (I_{CBO}/I_C) \qquad 0 < I_C < I_{PEAK} \qquad (739.7A)$$

$$\approx 1 - (I_{CBO}/I_C) - 1/[1 + \sqrt{(I_{Cit}/I_C}] \qquad I_{PEAK} \le I_C < I_{VALLEY} \qquad (739.7B)$$

$$\approx 1 - 1/[1 + \sqrt{(I_{Cit}/I_C)}] \qquad I_{PEAK} \ll I_C < I_{VALLEY} \qquad (739.7C)$$

$$\approx 1 - 1/[1 + \sqrt{(I_{Cit}/I_C)} + (I_C/I_{CHL})] \qquad I_{VALLEY} \le I_C \qquad (739.7D)$$

$$\approx 1 - 1/[1 + (I_C/I_{CHL})] \qquad I_{VALLEY} \ll I_C \qquad (739.7E)$$

The **off** or **blocking range** is given by (739.7A) and the characteristic is nearly the same as the leakage current with a slight multiplication, $I_C \approx MI_{CBO}$. Near the **peak voltage**, multiplication of the injected current is high and (739.7B) must be used to account for the increasing emitter efficiency or decreasing percentage of recombination in the emitter space-charge layer. The **negative resistance range** is almost completely controlled by the rapidly increasing emitter efficiency. It is given by (739.7D) which is in essence $\alpha_F M_p = 1$ or $\gamma_E M_p = 1$. At the **valley voltage**, the increasing emitter injection efficiency due to decreasing emitter space-charge-layer recombination is compensated by decreasing emitter injection efficiency due to increasing base injection level. In the valley voltage range, (739.7D) must be used to take into account these two opposing effects. The low differential on-resistance or the conducting range is mainly due to increasing base injection level or conductivity modulation and can be approximated by (739.7E).

The peak and valley voltages and currents can be obtained by setting

$$dV_{CB}/dI_C = 0$$

in (739.7B) and (739.7D). They are given by the following equations which relate the terminal currents and voltages to the transistor parameters. We ignore G_B/G_E and $X_B{}^2/2D_B\tau_B$ in $\alpha_B = \mathrm{sech}(X_B/L_B) \approx 1-(X_B{}^2/2D_B\tau_B)$ compared with 1 which is valid in BJT with high β_F(Max) or h_{FE}(Max) such as > 100. A limitation of the following result near the valley voltage and current is the omission of a collector and emitter series resistance which will invariably determine and significantly reduce the valley current. However, the valley voltage is not affected unless the series resistances are large.

$$I_{PEAK} \approx (4I_{CBO}{}^2 I_{EB})^{1/3} \tag{739.8A}$$

$$(V_{PEAK}/BV_{CBO})^n \approx 1 - (I_{CBO}/4I_{EB})^{1/3} - [1+(I_{EB}/2I_{CBO})^{1/3}]^{-1} \tag{739.8B}$$

$$I_{VALLEY} \approx (I_{EB} I_{CHL}{}^2/4)^{1/3} \tag{739.8C}$$

$$(V_{VALLEY}/BV_{CBO})^n \approx 1 - [1+(I_{EB}/2I_{CHL})^{1/3}+(I_{EB}/4I_{CHL})^{1/3}]^{-1} \tag{739.8D}$$

where $I_{EB} \equiv I_{Clt}$ is the emitter-base junction injection threshold.

These terminal currents and voltages can also be related to the current gain of the transistor to bring out a simple result and its physics. Using the common emitter forward current gain, $\beta_F = \alpha_F/(1-\alpha_F) \equiv h_{FE}$ and $\alpha_F = \alpha_B/[\ldots]$ in the starting equation, (739.7), we have

$$\begin{aligned} (V_{CB}/BV_{CBO})^n &= 1 - \alpha_F - (I_{CBO}/I_C) \\ &= (\beta_F+1)^{-1} - (I_{CBO}/I_C). \end{aligned} \tag{739.9}$$

Thus, at the peak voltage, we have

$$(V_{PEAK}/BV_{CBO})^n = [\beta_F(I_{PEAK})+1]^{-1} - (I_{CBO}/I_{PEAK}) \tag{739.10}$$

and at the valley voltage, we have

$$(V_{VALLEY}/BV_{CBO})^n = [\beta_F(I_{VALLEY})+1]^{-1} - (I_{CBO}/I_{VALLEY}) \tag{739.11}$$

$$\approx [\beta_F(I_{VALLEY})+1]^{-1} \approx 1/\beta_F \tag{739.11A}$$

or

$$V_{VALLEY} \approx BV_{CBO}/\beta_F{}^{1/n} = LV_{CEO}. \tag{739.12}$$

The valley result given above is very important in practice. It is the maximum V_{CE} one can operate a BJT in the common emitter configuration to avoid inadvertently biasing the BJT into the negative resistance region by switching transients or noise. In the negative resistance region the circuit may become unstable and oscillate (frequently seen on a transistor curve tracer), or the d.c. current may become very large to burn out the transistor if the series load resistance is small. This maximum voltage, V_{VALLEY}, is substantially smaller than the breakdown voltage of the collector-base junction in high gain or high β_F transistors as indicated by (739.12). For example, if $\beta_F=400$ and $n=6$, the maximum voltage is $\beta_F{}^{1/n}=400^{1/6}=2.7$ times smaller than BV_{CBO}. It is often the parameter given in BJT specification data sheets.

An approximation was made for the valley range to obtain (739.11A) from (739.11), because the valley current is much higher than the collector-base junction leakage current, I_{CBO}. However, this approximation may not be valid in transistors with low beta at low currents. Since I_{CBO} could vary widely (by one or more orders of magnitude) among production batches, V_{PEAK} is seldom given in the BJT data sheet. Another reason for its absence is that V_{PEAK} is larger than the valley voltage, V_{VALLEY}, and one would not want to operate the BJT above V_{VALLEY} to avoid oscillation and destructive thermal runaway.

Numerical Example

These theoretical results fit the experimental data such as the I_C vs $V_{CE}(I_B=0)$ of the 2N718 silicon n/p/n BJT in Fig.739.1. Assuming typical values of: $J_{CB0}=4.8\times10^{-10}$, $J_{Cit}=J_{EB}=8\times10^{-8}$, and $J_{CHL}=16$ A/cm^2 (which is low) then,

$$\begin{aligned}
J_{PEAK} &= 4.19\times10^{-9} \quad \text{A/cm}^2 \\
(V_{PEAK}/BV_{CBO})^n &= 0.699236 \\
J_{VALLEY} &= 1.73555\times10^{-2} \quad \text{A/cm}^2 \\
(V_{VALLEY}/BV_{CBO})^n &= 0.002428514 \\
(V_{PEAK}/V_{VALLEY})^n &= 287.9275
\end{aligned}$$

To compare with experimental data from the manufacturer's data sheet, we list in Table 739.1 the typical values of the negative resistance parameters and the difference in symbols used in the data sheet and in our analyses.

**Table 739.1
Example of Terminal Parameters of
Si n/p/n BJT (2N718)**

MANUFACTURERS' DATA SHEET		2N718	TEST AT	OURS
BV_{CBO}	Collector-base breakdown voltage	>100V	$V_{EB}=0$ I_C=Pulsed	BV_{CBO}
BV_{CEO}	Collector-emitter breakdown voltage	~100V	$I_B=0$	V_{PEAK}
LV_{CEO}	Collector-emitter sustaining voltage	41V	$I_B=0$	V_{VALLEY}
V_{CER}	Collector-emitter sustaining voltage	20V	R_BI_B=Pulsed $I_C=I_{VALLEY}$	V_{VALLEY}

We can fit these normalized results per unit area to the values given in the data sheet of the 2N718 n/p/n bipolar junction transistor to get an indication of the numbers. Since BV_{CBO} is usually not given accurately due to the high electric field at the surface perimeter of the collector-base junction, and V_{PEAK} is usually not measured due to instrumentation limitation in the factory, we use the following approach. An inspection of Fig.739.1 suggests a BV_{CEO} or V_{PEAK} of > 50V but it occurs in the pA range which is not shown on the mA scale for I_C of this figure and which is usually not given in the data sheet. From the starting p-type silicon used for these n/p/n transistors, the bulk breakdown voltage of the collector-base junction, BV_{CBO}, should exceed 100V. To estimate, we assume $V_{PEAK} \approx 100$V. Then, the impact multiplication exponent, n, is

$$n = \log_e(287.9275)/\log_e(V_{PEAK}/V_{VALLEY})$$

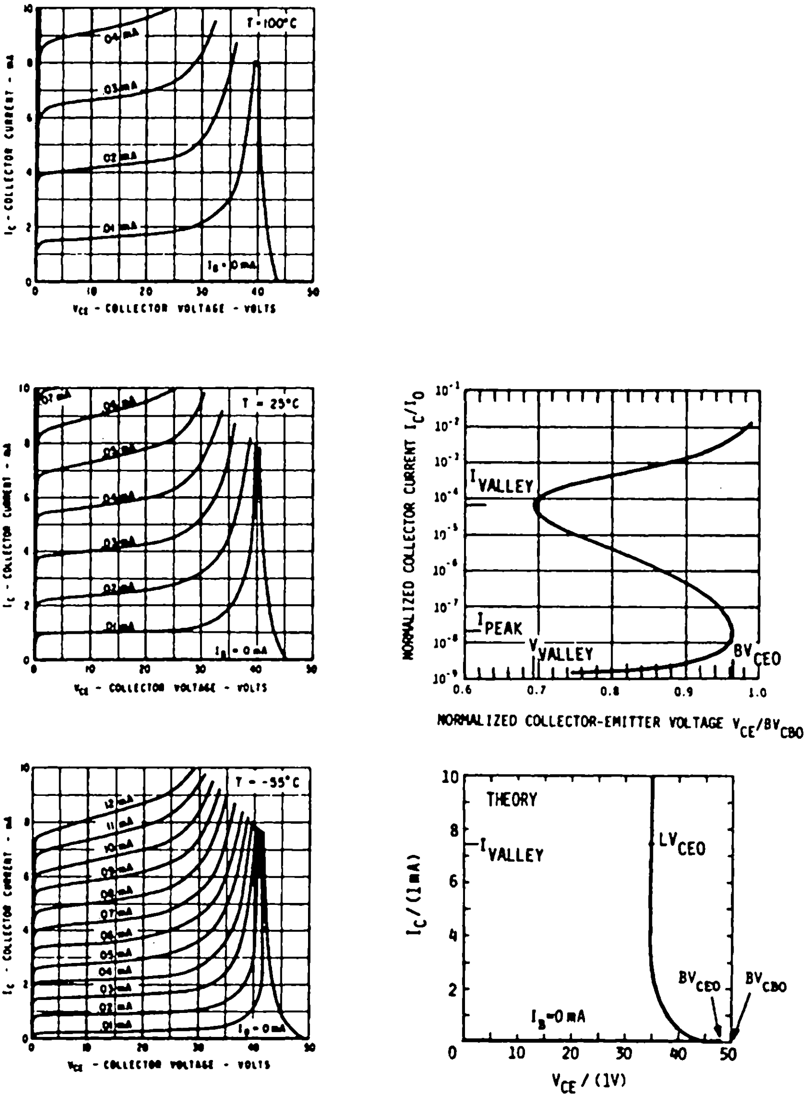
$$= \log_e(287.9275)/\log_e(100/41) = 6.351179$$

and the computed breakdown collector breakdown voltage is

$$BV_{CBO} = V_{PEAK}/0.699236^{1/n} = 100/0.945226 = 105.79\text{V}.$$

These computed values of n (for impact multiplication) and collector-base junction breakdown voltage are consistent with the expected values just discussed.

The theoretical I_C-V_{CE} curve using the above data and (739.7) is plotted in Fig.739.1(b) in both compact [lower figure (b)] and expanded [upper figure (b)] current scales. It is evident that the general shape of the experimental data at 25C and $I_B=0$ shown in Fig.739.1(a) agree with the compact theoretical curve shown in Fig.739.1(b) which was computed at $I_B=0$ or open base.



(a) (b)

Fig.739.1 Comparison of experimental and theoretical negative resistance current-voltage characteristics, I_C vs V_{CE} of an n/p/n Si BJT at $I_B=0$. (a) 2N718 and (b) normalized theory.

740 SMALL-SIGNAL CHARACTERISTICS OF BJT

In small-signal linear applications, the BJT is usually used as an amplifier in either wide band or tuned circuits. High input and output impedances, high current, voltage and power gains, and large bandwidth are desirable. These are the small-signal characterization parameters of the small-signal equivalent circuits which are analyzed to maximize the performance. Because of the desire for high gain and large bandwidth, the **common-emitter small-signal (CEss) equivalent circuit** is the most frequently taught in classroom and used in practice, while the **common-base small-signal (CBss) equivalent circuit** can only be found in books and articles of the 1950's to 1970's. The latest (1990) international competitions (disclosed at the annual International Electron Devices Meeting) on making the fastest homojunction BJTs in Si, and heterojunction BJTs in amorphous-Si/crystalline-Si/crystalline-Si, $Ga_xAl_yAs_z$, and $Si/Ge_xSi_{1-x}/Si$ layers are judged and bench-marked by the experimental small-signal f_t (frequency at unity short-circuit-output CE current gain, $\beta_f=1$), f_{Max} (maximum frequency of oscillation) and switching delay times. Switching delays will be analyzed in sections 75n.

However, the **common-base small-signal equivalent circuit** is the easiest to derive from the geometrical origin of each circuit element based on simple p/n junction device physics which we have already studied in chapter 5. It was also the earliest given after the BJT was invented in 1948 owing to the simple device physics of the geometrical origin. The common-emitter equivalent circuit was then derived from that of the common-base. We follow the physics path and develop the small-signal equivalent circuit of the common-base configuration first (section 741) and the formula of the cutoff frequency, f_α, followed by an application in which two figure-of-merits are derived, the f_{Max} and f_t. (section 742). Then, the common-emitter configuration is described (section 743) to define graphically and illustrate numerically the characteristic frequencies, f_β, f_t, and f_{bw} (section 744). First, the small-signal condition and the small-signal circuit analysis method are described in the following paragraphs.

The Small-Signal Condition

The small-signal condition in a multi-terminal (≥ 3) electron device under test (**DUT**) is defined as follows. **Small-signal** means that the amplitude of the time-dependent **applied signal** voltage (or current) to a pair of terminals (or a terminal pair to be designated as the **input terminal pair**) is so small that the input current (or voltage) **response** and the current and voltage **responses** at all other **output** terminal pairs are **proportional (linear functions)** of the input signal amplitude. Thus, the small-signal condition in a DUT is satisfied when the outputs faithfully reproduce a sinusoidal small-signal input without amplitude distortion and frequency multiplication and with only a phase delay. The same criterion applies if the input is a small-signal pulse. The output waveform will differ from the input but the amplitude of each harmonic frequency component of the input pulse will be amplified by the DUT faithfully and only the phase delay of each harmonic is different which causes the output waveform to be different from that of the input.

The major nonlinearity of a diode or a transistor is the exponential voltage dependence from the Boltzmann factors, $\exp(qV_{EB}/kT)$ and $\exp(qV_{CB}/kT)$, which appear as $[\exp(qV_{EB}/kT) - 1]$ and $[\exp(qV_{CB}/kT) - 1]$ as indicated in the Ebers-Moll equations. Since the collector-base junction is reverse biased, $\exp(qV_{CB}/kT)$ drops out. The emitter-base junction is forward biased, hence, the dominant nonlinear term is $\exp(qV_{EB}/kT)$. Thus, $[\exp(qV_{EB}/kT) - 1]$ must be linearized or made proportional to the small change in the voltage, V_{EB}, to study the small-signal properties. Section 540 discussed how small V_{EB} must be for analyzing the small-signal characteristics of p/n junction diodes. The derivation is repeated here. The criteria can be obtained by making a power series or Taylor series expansion of the exponential function exp(Z). Z can be a complex quantity. The expansion is

$$\exp(Z) = 1 + Z + Z^2/2! + Z^3/3! + \ldots. \quad (740.1)$$

or

$$\exp(Z) - 1 = + Z + Z^2/2! + Z^3/3! + \ldots. \quad (740.2)$$

$$= + Z\cdot(1 + Z/2) + Z^3/3! + \ldots. \quad (740.2A)$$

It shows that to linearize, the Z^2 and higher order terms must be dropped because they are nonlinear. To demonstrate nonlinearity, consider the Z^2 term. If Z is the INPUT voltage, then OUTPUT = AZ^2 = $A\cdot(\text{INPUT})^2$ where A is a constant. Thus, the output would increase parabolically or quadratically with the input rather than linearly. To maintain linearity at a specified accuracy, we must make the first nonlinear term, $Z^2/2$, much smaller than the linear term, Z, and then hope that the series behaves properly (i.e., converges) so that the sum of the higher order terms are even smaller than the first nonlinear term. Applying this criteria to (740.2A) gives $Z/2 << 1$. The fractional error is then given by Z/2. Engineering design requires a quantitative criterion. Thus, consider a maximum error or nonlinearity of 1% (or 1% distortion in a stereo amplifier), then we need to have $Z/2 \leq 0.01$ or $Z \leq 0.02$. This criteria can be used to determine the maximum allowable temporal variation in V_{EB} in order to have a valid small-signal or linear approximation. To demonstrate, we change V_{EB} by a small amount denoted by ΔV_{EB} and determine the maximum allowable ΔV_{EB}. Using the Taylor series expansion given above,

$$\exp[q(V_{EB}+\Delta V_{EB})/kT] = \exp(qV_{EB}/kT)\cdot[\exp(q\Delta V_{EB}/kT)] \quad (740.3)$$

$$= \exp(qV_{EB}/kT)\cdot[1 + (q\Delta V_{EB}/kT) + (q\Delta V_{EB}/kT)^2/2! + \ldots] \quad (740.4)$$

$$\approx \exp(qV_{EB}/kT)\cdot[1 + (q\Delta V_{EB}/kT)] \quad (740.4A)$$

The quadratic term in (740.4), $(q\Delta V_{EB}/kT)^2/2!$, is dropped compared with the linear term, $(q\Delta V_{EB}/kT)$, to linearize. This will be a good approximation if $(q\Delta V_{EB}/kT)^2 << 1$. For example, requiring 1% or better accuracy, we must have $(q\Delta V_{EB}/kT)/2 < 1/100$, i.e., a variation of V_{EB} or ΔV_{EB} smaller than

$$q\Delta V_{EB}/kT < 2/100$$

or

$$\Delta V_{EB} < (kT/q)/50 \approx 25/50 = 0.5 \text{ mV} \qquad (740.5)$$

where kT/q=25 mV at room temperature is used. This is a rather small voltage variation. It is an universal limit to maintain linear electrical conduction (current proportional to voltage) in heterogeneous solids where the local macroscopic particle density is exponentially dependent on the local macroscopic potential via the Boltzmann factor. [Quantum-mechanical tunneling has an even smaller limit and some conduction mechanisms have no linear limit because the lowest order response is proportional to $(\text{excitation})^n$ where $n \neq 1$.] This Boltzmann result gives a very important consequence: in order to have an accurate linearization, it is insufficient to have a small potential change compared with its steady-state or d.c. value such as $\Delta V_{EB} << V_{EB}$. The more stringent condition, $\Delta V_{EB} << 2kT/q \approx$ 50 mV, must be satisfied. For example, if V_{EB}=1.0V=1000mV, a 1% change in V_{EB} would give a ΔV_{EB} of 1000mV/100=10mV which is 10/0.5=20 times larger than the maximum limit on ΔV_{EB} (0.5 mV) to have a linearization accuracy of 1%.

Comparing the Charge-Control and Exact Small-Signal Analyses

There are two methods to derive the small-signal equivalent circuits, the exact small-signal expansion method and the charge-control method. We shall outline the procedures and underlying physics of these two methods in the next several paragraphs. The charge control method is then used in deriving the small-signal equivalent circuit models of the BJTs in the CB and CE configurations.

Exact Analysis for Small-Signal Equivalent Circuit Models

The **exact method** uses the linear expansion $x_A(t) = X_A + x_a(t) = X_A + X_a(\omega)\exp(j\omega t)$ for the sinusoidal steady-state case. The variable X_A is the d.c. or time-average value of the variable $x_A(t)$ at node A. $X_a(\omega)$ is the complex rms (root-mean-square) value at the signal frequency ω that can be measured with a voltmeter, or the peak value of the waveform that can be displayed on an oscilloscope to show the peak value and the phase shift. For example, in the continuity equation for holes (350.2), the time derivative $\partial p(x,t)/\partial t$ at a node labeled x or a position x, can be worked out using $p(x,t) = P(X) + p(x,\omega)\exp(j\omega t)$,

$$\partial p(x,t)/\partial t = (\partial/\partial t)[P(X) + p(x,\omega)\exp(j\omega t)] \qquad (740.6)$$

$$= j\omega p(x,\omega)\exp(j\omega t) \qquad (740.6A)$$

$$= j\omega[p(x,t) - P(X)] \qquad (740.6B)$$

$$= j\omega\Delta p(x,t) \qquad (740.6C)$$

This term is added to the generation-recombination term in the continuity equation (350.2), $q(g_P - r_P)$. Using the constant lifetime defined by (370.2) and worked out for τ_n in the SRH model given by (372.1D), then the two terms in the hole continuity equation (350.2) that gives the divergence of the hole current density can be combined to give two 1-d continuity equations, one for the d.c. hole concentration, P(X) or $\Delta P(X) = P(X) - P_{Equilibrium}$, and the other for the small-signal time-varying hole concentration, $\Delta p(x,t) = p(x,t) - P(X)$. Thus,

$$- dj_P(x,t)/dx \equiv - (d/dx)[J_P(X) + j_p(x,t)] \tag{740.7A}$$

$$= \partial p(x,t)/\partial t - q(g_P - r_P) \tag{350.2}$$

$$= j\omega\Delta p(x,t) + \Delta p(x,t)/\tau_p + \Delta P(x)/\tau_p. \tag{740.7B}$$

So the d.c. and small-signal equations are

$$- dJ_P(x)/dx = \Delta P(x)/\tau_p \tag{740.8A}$$

and

$$\begin{aligned} - dj_p(x,t)/dx &= \Delta p(x,t)\cdot[(1/\tau_p) + j\omega] \\ &= \Delta p(x,t)/[\tau_p/(1+j\omega\tau_p)] \\ &\equiv \Delta p(x,t)/\tau_{p'}, \end{aligned} \tag{740.8B}$$

where

$$\tau_{p'} = \tau_p/(1+j\omega\tau_p). \tag{740.8C}$$

Note that (740.8A) and (740.8B) are identical in form if τ_p in the d.c. equation (740.8A) is replaced by $\tau_{p'} = \tau_p/(1+j\omega\tau_p)$. Thus, **the d.c. solutions obtained for** $\Delta P(x) \equiv P(x) - P_{Equilibrium}$ from solving the d.c. continuity equation (740.8A) can be used to give the small-signal solutions of (740.8B) if τ_p in the d.c. solutions is replaced by $\tau_{p'}$ or $\tau_p/(1+j\omega\tau_p)$. This powerful shortcut eliminates all the additional and tedious algebra of obtaining the exact small-signal solutions of the diffusion and linearized recombination-generation currents in diodes and transistors. We shall freely use this result to compare the approximate solutions with the exact solutions we shall cite using this shortcut.

The exact small-signal analysis requires the simultaneous solution of the six small-signal **partial differential equations** from the six general Shockley equations using the solution of the six d.c. partial (2-d or 3-d but becomes ordinary if 1-d) differential equations. The above small-signal hole continuity equation is just one of the six which gives explicit solution for minority carrier transport. A systematic procedure has been developed for more general solutions by this author which solves these simultaneous partial differential equations using the elementary circuit laws (Kirchoff's laws) and the well-developed and canned matrix algebra algorithms. It is known as the **Circuit Technique for Semiconductor Analysis (CTSA)**. A brief outline of this method is now given. The first step of CTSA is to make a small-signal expansion of the current, voltage, and voltage- or current-dependent material variables of the six Shockley equations. This expansion transforms the six

Shockley equations into two sets of partial differential equations: six d.c. equations and six small-signal equations. In the second step, the six small-signal **partial differential equations** are transformed into six small-signal **difference equations**. Similarly, the six d.c. partial differential equations are transformed into six difference equations. Finally, the small-signal equivalent circuit is **synthesized** from the six small-signal difference equations and the d.c. equivalent circuit is synthesized from the six d.c. difference equations. The variables of the small-signal equivalent circuit are small-signal electron and hole conduction currents (including both drift and diffusion currents), the displacement current, and the small-signal quasi-Fermi potentials of the electrons, holes, and trapped electrons for an electron trap type of generation-recombination-trapping center (or trapped holes for a hole trap), and the small-signal electric potential. The variables of the d.c. equivalent circuit are the electron and hole conduction currents (including both drift and diffusion), and the quasi-Fermi potentials of the electrons, holes, and trapped electrons (or trapped holes), and the electrostatic potential. The derivation algebra via analysis and synthesis is simple and straightforward. The resultant small-signal and d.c. equivalent circuits are three-dimensional for the general case. In the one-dimensional case, the two equivalent circuits reduce to two **transmission lines**. The small-signal circuit is a **three-wire transmission line** consisting of the **hole conduction current line** (with hole quasi-Fermi potential at each x on the line), **electron conduction current line** (electron quasi-Fermi potential), and **displacement current line** (electric potential). The adjacent trapped electron quasi-Fermi potentials, at x and $x+\Delta x$, are not interconnected by any element unless there is intertrap tunneling or the trapping centers can diffuse and drift like that in biomemberances in which case it would be represented by an **ion current line** from the migration of electron-hole traps. The d.c. circuit is a **two-wire transmission line** consisting of the **hole and electron conduction currents**. The series elements are the hole and electron conductances and the dielectric capacitance. The shunt element at each node, X, is a T-section consisting of the electron and hole trapping conductances and trapped charge storage capacitance. The development of the equivalent circuit began in earnest in 1958 by Shockley who first derived the small-signal equations at thermal equilibrium for just one species of SRH recombination-generation-trapping centers [740.1]. The complete exact small-signal equivalent circuit at **any arbitrary steady-state or non-equilibrium conditions**, and the d.c. and large-signal circuits were then derived and developed to completion in 1961 by Sah. Applications of the CTSA model in the following two decades were recently reviewed [740.2] which gives references where the details of the transmission line and distributed models can be found.

[740.1] W. Shockley, "Electrons, holes and traps," Proc.IEEE, 46(6), 973-990, June 1958.

[740.2] C. T. Sah, "New integral representations of circuit models and elements for the circuit techniques for semiconductor device analysis," Solid State Electronics, 30(12), 1277-1281, Dec. 1987.

Charge-Control Analysis for Small-Signal Equivalent Circuit Models

The **second method** is an approximate method using the charge-control approach. The small-signal resistances and conductances are equated to the differential resistance and conductance,

$$r = dV/dI = \lim_{\Delta I \to 0}(\Delta V/\Delta I) \tag{740.9A}$$

and

$$g = dI/dV = \lim_{\Delta V \to 0}(\Delta I/\Delta V). \tag{740.9B}$$

The small-signal charge storage capacitance is equated to

$$C = dQ/dV = \lim_{\Delta V \to 0}(\Delta Q/\Delta V). \tag{740.9C}$$

These are **exact** by definition. But they are **not** the values measured by a sinusoidal small-signal impedance bridge (the so-called a.c. bridges). The bridge applies a small sinusoidal test signal to the DUT and measures the rms magnitude and the phase of the response. The charge-control definition requires a measurement of the total response after a small test signal in the form of a small voltage or current step is applied to the DUT. For example, the conductance is computed from the measured total current change, $\Delta I = i(t=\infty) - i(0)$, after a voltage step, $\Delta V = v(t=0^+) - v(t=0^-)$ is applied to the DUT at $t=0$. Similarly, the capacitance is computed from the measured total charge change, $\Delta Q = q(t=\infty) - q(0) = \int i(t)dt$, after a voltage step is applied to the DUT at $t=0$. Obviously, it requires an integrating current meter to measure the total charge flowing into the entire device layer through a particular terminal. Thus, the charge-control model lumps the charges distributed spatially in the device structure; hence, it is known as the **single-lump**, **one-lump**, or just the **lumped model**. It is not an approximation in itself because it is derived from precise definitions and in fact, it also has an exact experimental base. But it is an approximation to the commonly measured small-signal resistance and capacitance using an a.c. admittance bridge.

The fundamental difference between the exact and charge-control methods is obvious. The charge control method measures the **small-signal transient response**, not the time-development of the response but the total response, while the exact method or the a.c. impedance bridge measures the **small-signal sinusoidal steady-state response**. The resulting difference is that the charge-control conductances and capacitances do not take into account the propagation delay of the small-signal passing through the layers of the DUT from the originating node or input terminal to the distance node or output terminal. In a BJT, these are the five layers already used in the d.c. analysis in section 733 and shown in Fig.733.1(c): the emitter quasi-neutral, emitter-base space-charge, base quasi-neutral, base-collector space-charge, and collector quasi-neutral layers. For example, in the p/n/p BJT, the transmission of a small signal from the emitter terminal to the collector terminal is carried by the small variation of the density of the holes, $p(x,t)=p(X,t)-P(X)=\delta p(X,t)$, injected by the forward voltage applied to the emitter-base terminals and collected by the reverse biased collector-base junction. And it takes time for the holes to pass through the layers, by diffusion and drift. But, the charge-control model does not measure this time delay, rather it measures the total change after a long time from the instance of applying a change at the input. If the one lump is divided into many lumps, then the charge control solution would approach the distributed or exact solution. Fortunately, the emitter and base layers through which the minority carriers have to pass are so thin in high gain BJTs that the charge-control method gives an accurate numerical solution with small errors which can be corrected by numerical multipliers of the order of unity.

Nevertheless, the charge-control method will be used to analyze the small-signal properties of BJTs. It was also used to analyze the MOS capacitor in sections 41n, the p/n junction diode in section 54n, and the MOS transistor in sections 65n. The reason is that the algebra of the exact method is more tedious and could obscure the important physics for a beginner, while the charge-control method focuses and identifies the circuit elements with the fundamental drift, diffusion, recombination-generation-trapping mechanisms. However, accurate models and numerical results of the characteristic frequencies of BJTs, obtainable only from the exact analyses using either the differential equations or the CTSA from the corresponding difference equations, are becoming increasingly important. They are necessary to guide and gauge the geometry and structure designs of the ultrafast BJTs and to guide the manufacturing of the ultrafast integrated circuits because the recent technology advancements are making it possible to fabricate the nearly ideal BJT (having only the intrinsic part and no parasitics). The correctional numeric multipliers will be used to make the charge-control solutions more accurate when the numerical illustration examples are computed.

741 Common-Base Small-Signal Tee (CBss-Tee) Models of BJT

The Common-Base small-signal Tee (CBss-Tee) equivalent circuit is derived first for four reasons although the common-emitter small-signal hybrid-pi (CEss-Hπ) model is preferred in applications: (1) its derivation is most straightforward, (2) it is historically the first model of BJT, (3) it gives a direct and swift derivation of the small-signal figure-of-merit of BJT, the maximum frequency of oscillation, which has been used as the benchmark to compare the state-of-the-art of ultrafast and ultrahigh frequency BJT technologies, and (4) the common-emitter hybrid-pi model can be directly derived from the common-base tee model. The CBss-Tee equivalent circuits for the ideal intrinsic transistor are shown in Figs.741.1(b)-(d) and the parasitics are added in Fig.741.1(e) which are now derived.

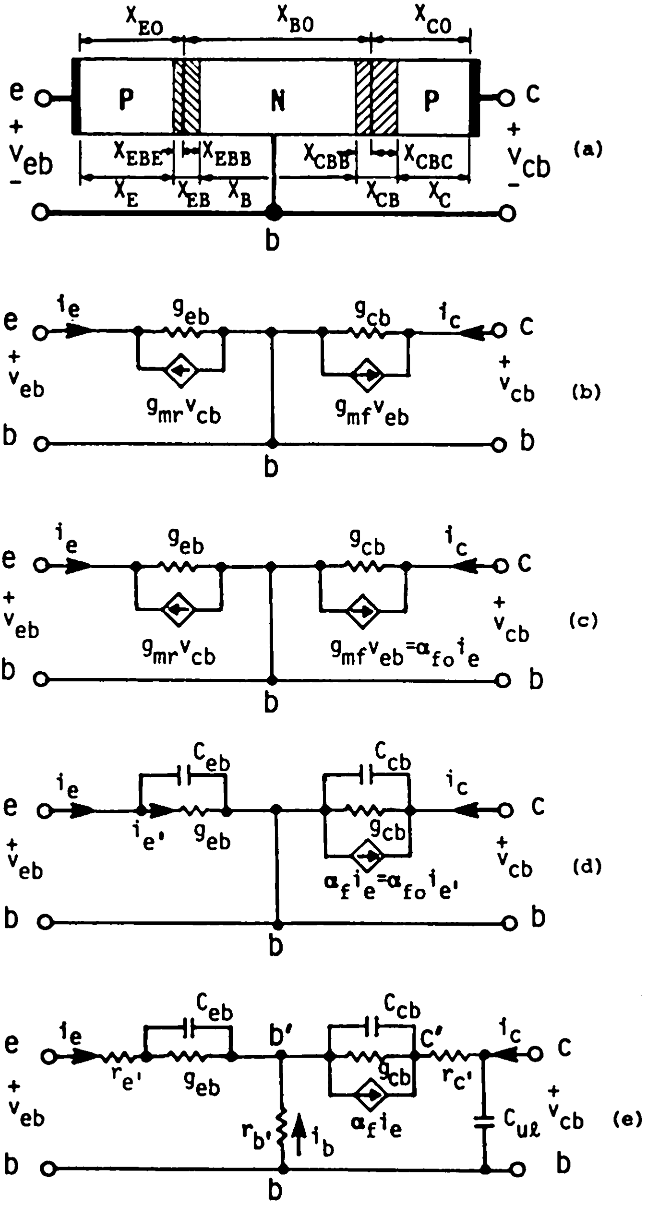

X_{EO}
X_{BO}
X_{CO}
e
P
N
P
C
+
V_{eb}
-
X_{EBE}
X_{EBB}
X_{CBB}
X_{CBC}
+
V_{cb}
-
X_E
X_{EB}
X_B
X_{CB}
X_C
b
(a)
e
i_e
g_{eb}
g_{cb}
i_c
C
+
v_{eb}
$g_{mr}v_{cb}$
$g_{mf}v_{eb}$
+
v_{cb}
b
b
b
(b)
e
i_e
g_{eb}
g_{cb}
i_c
C
+
v_{eb}
$g_{mr}v_{cb}$
$g_{mf}v_{eb}=\alpha_{fo}i_e$
+
v_{cb}
b
b
b
(c)
C_{eb}
C_{cb}
e
i_e
$i_{e'}$
g_{eb}
g_{cb}
i_c
C
+
v_{eb}
$\alpha_f i_e=\alpha_{fo}i_{e'}$
+
v_{cb}
b
b
b
(d)
C_{eb}
C_{cb}
e
i_e
$r_{e'}$
g_{eb}
b'
c'
g_{cb}
$r_{c'}$
i_c
C
+
v_{eb}
$r_{b'}$
i_b
$\alpha_f i_e$
$C_{u\ell}$
+
v_{cb}
b
b
b
(e)

Fig.741.1 The common-base small-signal tee (CBss-Tee) equivalent circuits of an intrinsic BJT with parasitics added in part (e). (a) The cross-sectional view of the 5-layer model. (b) D.C. and low-frequency. (c) Simplified and derived D.C. and low-frequency. (d) D.C. to high-frequency for intrinsic BJT. (e) D.C. to high frequency with parasitics, $r_{e'}$, $r_{b'}$, $r_{c'}$, and $C_{u\ell}$ added.

The CBss-Tee equivalent circuit model is known as the **short-circuit admittance** (or y-parameter) **two-port network** represented by the pair of two-port network equations

$$I_1 = y_{11}V_1 + y_{12}V_2 \qquad (741.1A)$$

$$I_2 = y_{21}V_1 + y_{22}V_2. \qquad (741.1B)$$

The input node is labeled 1 (emitter terminal in the CB configuration). The output node is labeled 2 (collector terminal in the CB configuration). The reference node is the common node and not labeled (base terminal in the CB configuration). The currents are flowing into the nodes from the external lead. It is evident from Fig.741.1(b) that $g_{eb} = g_{11}$ (the real part of $y_{11} = g_{11} + jb_{11}$), $g_{mr} = -g_{12} > 0$, $g_{mf} = -g_{21} > 0$, $g_{cb} = g_{22}$, v_{eb} or $V_{eb} = V_1$, v_{cb} or $V_{cb} = V_2$, i_e or $I_e = I_1$, and i_c or $I_c = I_2$. The direction of the two dependent current sources are reversed so that the transconductances, g_{mf} and g_{mr} are positive and real quantities.

To derive the formulae of the two-port short-circuit conductance and susceptance or admittance parameters of CBss-Tee equivalent circuit model of BJT in terms of device geometry and material properties, we shall: (i) take the differential of the two d.c. current equations to obtain the conductance elements; (ii) use the charge control method to obtain capacitance elements; and (iii) consider the diffusion delay in the base to obtain the frequency dependence of the current amplification factor. The results are then compared with the exact solutions.

Low Frequency CBss-Tee Model of Intrinsic BJT

The general low-frequency CBss-Tee equivalent circuit of the intrinsic BJT is shown in Fig.741.1(b). Intrinsic means that the device is one-dimensional and ideal and that the parasitics are excluded. The formulae of the low-frequency equivalent circuit elements (all resistive) are obtained by taking the differential of the two d.c. transistor current equations given by (733.19) and (733.20) which are repeated below in (741.2A) and (741.2B).

$$J_E = (j_E + j_{EB} + j_B)[\exp(qV_{EB}/kT) - 1] - \alpha_B j_B[\exp(qV_{CB}/kT) - 1] \qquad (741.2A)$$

$$J_C = (j_C + j_{CB} + j_B)[\exp(qV_{CB}/kT) - 1] - \alpha_B j_B[\exp(qV_{EB}/kT) - 1]. \qquad (741.2B)$$

Denote the differential or increment, ΔA_A, by the small-signal variable a_a, then $\Delta I_E = i_e$, $\Delta V_{EB} = v_{eb}$, $\Delta I_C = i_c$ and $\Delta V_{CB} = v_{cb}$. Let the emitter junction area be A_E. Then, the two d.c. current density equations (741.2A) and (741.2B) can be multiplied by A_E to give the pair of small-signal **two-port short-circuit conductance equations** of the intrinsic BJT, which are

$$i_e = (\partial I_E/\partial V_{EB})\Big|_{V_{CB}} v_{eb} + (\partial I_E/\partial V_{CB})\Big|_{V_{EB}} v_{cb} \qquad (741.3)$$

and

$$i_c = (\partial I_C/\partial V_{EB})\Big|_{V_{CB}} v_{eb} + (\partial I_C/\partial V_{CB})\Big|_{V_{EB}} v_{cb}. \qquad (741.4)$$

Simplifying by denoting the partial derivatives as the self conductances and mutual conductances or transconductances connected between the two nodes designated by the subscript, we have

$$i_e = g_{eb}v_{eb} - g_{mr}v_{cb} = I_1 = g_{11}V_1 + g_{12}V_2 \qquad (741.5)$$

and

$$i_c = -g_{mf}v_{eb} + g_{cb}v_{cb} = I_2 = g_{21}V_1 + g_{22}V_2. \qquad (741.6)$$

The g_{xy}'s are known as the **short-circuit conductance parameters** in two-port network theory, i.e., they are the conductance looking into a terminal and its reference terminal with the other terminal short-circuited to the reference terminal. For example, g_{eb} (= g_{11} in two-port notation) is the conductance looking into the terminals e and b with the other two terminals, c and b, shorted. This is known as the emitter-base **short-circuit input conductance**. The CBss-Tee equivalent circuit given in Fig.741.1(b) can be immediately synthesized from the two current equations, (741.5) and (741.6), using this short-circuit concept.

The formulae of the conductances in terms of the device material and geometry properties can be readily obtained using (741.2A) and (741.2B), and the definitions of the current coefficients, j_E, j_{EB}, j_B, j_{BC}, j_C and α_B given in (733.21A) to (733.21F), and their approximate forms given in section 737. The input conductance is

$$\begin{aligned} g_{eb} = g_{11} &= (\partial I_E/\partial V_{EB})\Big|_{V_{CB}=\text{constant or } v_{cb}=0} \\ &= A_E(\partial/\partial V_{EB})\{(j_E+j_{EB}+j_B)[\exp(qV_{EB}/kT)-1]\} \\ &= g_e + g_{ebt} + g_b. \end{aligned} \qquad (741.7)$$

It contains three terms from the three layers of the emitter-base junction of the p/n/p BJT. The controlling transport mechanisms are: (1) minority carrier (electrons) diffusion-drift and recombination with the majority carrier (holes) in the quasi-neutral p-type emitter layer which gives g_e; (2) holes recombining with electrons at the SRH traps or centers in the emitter-base junction space-charge or transition layer which gives g_{ebt}, where the third subscript t signifies both the transition layer and the traps; and (3) minority carrier (holes) diffusion-drift and recombination with majority carriers (electrons) in the quasi-neutral n-type base layer which gives g_b. The third subscript, t, in g_{ebt} of (2) is needed because g_{eb} is used to denote the total conductance connected between the emitter and base nodes, e and b. The BJT is designed to have high power gain or nearly unity alpha and large beta, which results in $g_b >> g_{ebt}+g_e$. The formulae of g_b is obtained using the approximation for j_B given by (737.9A) and it is

$$g_b = A_E(q/kT)j_B\exp(qV_{EB}/kT)$$
$$= (q/kT)(I_C-I_{CBO})/\alpha_B \approx (qI_C/kT)/\alpha_B. \qquad (741.8)$$

The conductances g_e and g_{ebt} can be readily obtained in a similar way and are left as homework problems. [See P741.1 to P741.3.]

The **short-circuit output conductance** at the collector-base terminals, g_{cb} (g_{22} in two-port notation), contains four terms from two sources.

$$g_{cb} = g_{22} = (\partial j_C A_E/\partial V_{CB})\Big|_{V_{EB}}$$
$$= A_E(\partial/\partial V_{CB})\{(j_B+j_{BC}+j_C)[\exp(qV_{CB}/kT)-1] - \alpha_B j_B[\exp(qV_{EB}/kT)-1]\}$$
$$= \qquad g_{b'} + g_{cbt} + g_c \qquad + \ g_a. \qquad (741.9)$$

The first three terms are from one source, i.e., associated with the collector-base junction alone. They are due to (i) carrier generation-recombination in the collector-base junction space-charge or transition layer, g_{cbt}; and the diffusion-drift-recombination-generation, (ii) in the quasi-neutral base layer, $g_{b'}$, and (iii) in the quasi-neutral collector layer, g_c. Note $g_{b'} << g_b$ of (741.8) because $g_{b'}$ is the conductance of the leakage current of the reversely biased collector-base junction and $g_b \approx qI_E/kT$ is the large forward conductance of the forward biased emitter-base junction. The electron-hole generation conductance in the space-charge layer of a reverse biased collector-base junction, g_{cbt}, is the largest in Si and other large energy-gap semiconductors since generation rates in the two quasi-neutral layers are proportional to the minority carrier concentrations which drop exponentially to zero with reverse collector bias because they are proportional to $\exp(qV_{CB}/kT) = \exp(-|V_{CB}/kT|) << 1$.

The fourth term, g_a, comes from a second source, and is generally the dominant term. It arises from the **Early effect** discussed in section 738 owing to the modulation of the base layer thickness by the time varying small-signal across the collector-base junction. The subscript 'a' comes from Early since E or e is used to denote the emitter. Thus, g_a comes from the dependence of j_B and α_B on the base layer thickness which varies with time from $v_{CB}(t) = V_{CB}+v_{cb}(t)$. There are two contributing terms in (741.9), $j_B[\exp(qV_{CB}/kT)-1]$ and $\alpha_B j_B[\exp(qV_{EB}/kT)-1]$. Obviously the latter dominates because $V_{EB}>0$ and $V_{CB}<<0$. Usually, the electric base is thin, $X_B<<L_B$, in order to give high current transfer ratios, alpha and beta. As indicated in (737.2B), $\alpha_B = \mathrm{sech}(X_B/L_B) \approx 1 - X_B^2/(2D_B\tau_B) \approx 1$, and $j_B = qD_BP_B/X_B$ from (737.3) on Gummel number for the base. Thus, both α_B and j_B increases with decreasing X_B but the main effect comes from X_B^{-1} of j_B. X_B is indicated in Fig.741.1(a) and its dependence on V_{CB} can be illustrated using the results of the abrupt p/n junction given by (523.16) to (523.20) with constant donor impurity concentration in the base, N_{BB}, and constant acceptor impurity concentration in the collector, N_{CC}. This gives [See Fig.741.1(a)]

$$X_B = X_B(V_{CB}) = X_{B0} - X_{EBB} - X_{CBB}$$

$$= X_{B0} - X_{EBB} - [N_{CC}/(N_{BB}+N_{CC})]X_{CB}(V_{CB}) \qquad (741.9A)$$

where from (523.19) and (531.1),

$$X_{CB} = X_{CB}(V_{CB}) = \sqrt{[2\varepsilon_s N_{CC}N_{BB}(V_{CB}-V_{CB})/q(N_{CC}+N_{BB})]}. \qquad (741.9B)$$

X_{B0} is the fixed or time-invariant metallurgical or geometrical base layer thickness while X_B is the quasi-neutral base layer thickness or the **electrical base layer thickness**. Shown in Fig.741.1(a), $X_{CBB} = [N_{CC}/(N_{BB}+N_{CC})]X_{CB}$ is the part of the collector-base space-charge layer on the base side of the collector-base junction.

Instead of writing out all the terms explicitly, the Early conductance can be written in the following simplified and yet general form which immediately gives a feel of its magnitude

$$g_a = - A_E[\exp(qV_{EB}/kT)-1]\cdot[(\partial\alpha_B j_B/\partial V_{CB})] \qquad (741.10)$$

$$= - A_E[\exp(qV_{EB}/kT)-1]\cdot[(\alpha_B j_B/V_{CB})(\partial\log_e\alpha_B j_B/\partial\log_e V_{CB})]$$

$$= |(I_C-I_{CBS})/V_{CB}|\cdot E_A \qquad (741.10A)$$

$$= \qquad G_{CBS} \qquad \cdot E_A \qquad (741.10B)$$

$$= |(I_C-I_{CBS})/(|V_{CB}|-V_A)| \qquad (741.10C)$$

where

$$E_A = \partial\log_e|\alpha_B j_B|/\partial\log_e|V_{CB}| \approx \partial\log_e|j_B|/\partial\log_e|V_{CB}| \qquad (741.10D)$$

$$= \partial\log_e|X_B|/\partial\log_e|V_{CB}|$$

$$= [N_{CC}/(N_{BB}+N_{CC})]\cdot(-\partial\log_e|X_{BC}|/\partial\log_e|V_{CB}|) \qquad (741.10E)$$

$$= [N_{CC}/(N_{CC}+N_{BB})]\cdot m. \qquad (741.10F)$$

G_{CBS} is the d.c. conductance of the collector-base junction at a constant applied d.c. emitter-base voltage. It is proportional to the collector terminal d.c. current and hence has a very large value at high I_C, high I_E or high forward V_{EB}. E_A is the Early coefficient, consisting of two multiplying factors indicated in (741.10F). (i) $m=1/2$ is the slope of a plot of $\log X_{CB}$ versus $\log V_{CB}$ or the exponent in $X_{CB} = B|V_{CB}|^m$ for the abrupt n/p junction with constant dopant impurity concentration. (ii) The fraction $N_{CC}/(N_{CC}+N_{BB})$ is the fractional penetration of the collector-base junction space-charge layer into the base layer. The Early effect can be reduced if $N_{BB} >> N_{CC}$ to make this penetration into the base very small. The alloy BJTs manufactured during the 1950's had just the opposite condition that made the Early effect very large: N_{CC}(dopant in the alloy used to form the collector) $>>$ N_{BB}(substrate), so that this fraction was unity, $N_{CC}/(N_{CC}+N_{BB}) \approx N_{CC}/N_{CC} = 1$, i.e. all the collector-base junction space-charge layer laid in the base layer. For the present and future diffused and epitaxial junctions with graded impurity profile, the Early factor E_A cannot be readily separated into two terms. But it can be made as small as desirable by impurity profile design, such as using a lowly doped collector (LDC) drift layer in the p/n/i/p or n/p/i/n structure, so that the major fraction of the collector-base junction space-charge layer penetrates into the collector body or substrate, making the variation of X_B with $v_{cb}(t)$ very small and nearly zero. E_A can be directly computed from the forward (also reverse) output characteristics, $I_C(V_{EB}=\text{constant})$ versus V_{CB}, such as Fig.732.1(a') obtained from the product data sheet.

An Early voltage for the common-base configuration, V_{Acb}, can also be defined empirically as indicated by (741.10C) in analogy to the Early voltage V_A defined in (738.6A)-(738.7A) for the common emitter configuration. Note that if the d.c. emitter current is held constant instead of the d.c. emitter-base voltage, such as Fig.732.1(a), then g_A drops out exactly. [See Problem P741.5.]

The **reverse transconductance**, g_{mr} (or g_{12} in two-port notation), is normally negligible and can be ignored because the collector junction is reverse biased and hence, is not injecting any holes into the base layer so that no holes from the collector junction would reach the emitter-base junction to affect its voltage. However, if the Early effect is large, then a small g_{mr} may exist due to the leakage current multiplied by the Early effect. Thus, this feedback current source, $g_{mr}v_{cb}$, in the emitter branch in Fig.741.1(b) is retained in Figs.741.1(c) but it is dropped in Figs.741.1(d) and (e). Following the expansion given in (741.3) and the definition in (741.5), g_{mr} is

$$\begin{aligned}
g_{mr} &= -\,A_E(\partial J_E/\partial V_{CB})\Big|_{V_{EB}} = +\,A_E(\partial/\partial V_{CB})\cdot\{\alpha_B j_B[\exp(qV_{CB}/kT)-1]\}\\
&\simeq +\,A_E\alpha_B\cdot[\exp(qV_{CB}/kT)-1]\cdot(\partial j_B/\partial V_{CB})\\
&\simeq +\,A_E\alpha_B\cdot[\;0 \qquad\qquad -1]\cdot(\partial j_B/\partial V_{CB})\\
&= +\,A_E\alpha_B\cdot(-j_B/V_{CB})\cdot E_A = |I_{CBS}/V_{CB}|\cdot E_A. \qquad (741.11)
\end{aligned}$$

The **forward transconductance**, g_{mf} (or g_{21} in two-port notation), is the parameter that makes the BJT an amplifier or oscillator. It is given by

$$\begin{aligned}
g_{mf} &= -\,A_E(\partial J_C/\partial V_{EB})\Big|_{V_{CB}} = +\,A_E(\partial/\partial V_{EB})\{\alpha_B j_B[\exp(qV_{EB}/kT)-1]\}\\
&= +\,A_E\alpha_B j_B(q/kT)\exp(qV_{EB}/kT)\\
&\simeq q|I_C|/kT = \alpha_B g_b. \qquad (741.12)
\end{aligned}$$

This result immediately suggests that the dependent current source of the transconductance, $g_{mf}v_{eb}$, in the CBss-Tee of Fig.741.1(c) can be replaced by another dependent current source which depends on the input current, i_e, since

$$\begin{aligned} g_{mf}v_{eb} &= \alpha_B g_b v_{eb} = \alpha_B\cdot(g_b/g_{eb})\cdot(g_{eb}v_{eb}) \\ &= \alpha_B\cdot[g_b/(g_e+g_b+g_{ebt})]\cdot i_e \\ &= \alpha_B\gamma_{eo}i_e = \alpha_{fo}i_e \end{aligned} \tag{741.13}$$

where the **low-frequency small-signal alpha** is defined by

$$\alpha_{fo} = \alpha_B\gamma_{eo} \tag{741.13A}$$

and the **low-frequency small-signal emitter injection efficiency** is defined by

$$\gamma_{eo} = g_b/(g_{ebt}+g_b+g_e). \tag{741.13B}$$

The alternate equivalent circuit is also shown in Fig.741.1(d) with the alternative expression for the dependent transfer current source, $\alpha_{fo}i_e$. It is known as the **Common Base small-signal hybrid-Tee (CBss-hTee)** model. Note that this is just the small-signal expansion of the explicit d.c. CB hybrid-T model shown in Fig.735.1(b).

High-Frequency CBss-Tee Model of Intrinsic BJT

At high-frequencies, capacitances due to charges stored in the five layers of the BJT shown in Fig.741.1(a) must be taken into account since they delay the transistor's response to an input signal. The capacitances are now computed using the charge control method which also reveals the charging and discharging mechanisms and the origin of the carriers or mobile charges. To proceed, we write down the total charge stored in the entire p/n/p transistor as a sum of the charges stored in the five layers shown in Fig.741.1(a), the emitter, base, and collector quasi-neutral layers, and the emitter-base and collector-base junction space-charge layers. There are six terms, three each introduced by the applied emitter-base and collector-base junction voltages:

$$\begin{aligned} Q = &+ Q_E(V_{EB}) + Q_{EB}(V_{EB}) + Q_B(V_{EB}) \\ &+ Q_C(V_{CB}) + Q_{CB}(V_{CB}) + Q_B(V_{CB}). \end{aligned} \tag{741.14}$$

The junction voltage dependence of each charge term is explicitly indicated to facilitate the differentiation operation, d/dV, to give the capacitance. The charges in the three quasi-neutral layers, Q_E, Q_B, and Q_C are the minority carrier charges while Q_{EB} and Q_{CB} are majority carrier charges inside or at the edge of the space charge layers.

The emitter-base junction capacitance is given by $C_{eb} = \partial Q/\partial V_{EB}$ (V_{CB}=constant). It contains four terms similar to the four terms of a forward biased p/n junction with very thick quasi-neutral emitter and collector layers given by (541.8) to (541.12). However, the two quasi-neutral capacitances obtained for the p/n junction must be modified because the quasi-neutral emitter and base layers in a BJT are very thin compared with the minority carrier diffusion length. Thus,

$$C_{eb} = A_E(\partial Q/\partial V_{EB})\Big|_{V_{CB}=0} = A_E(d/dV_{EB})(Q_E + Q_{EB} + Q_B)\Big|_{V_{CB}=0} \qquad (741.15)$$

$$= C_e + C_{ebt} + C_b. \qquad (741.16)$$

The space-charge layer capacitance of the emitter-base junction contains two contributions: $dQ_{EB}/dV_{EB} = C_{ebt} = C_{eb\text{-}pn} + C_{eb\text{-}d}$. $C_{eb\text{-}pn}$ is the capacitance from electrons and holes injected into and stored inside the space-charge layer of the emitter-base junction. $C_{eb\text{-}d}$ is the capacitance due to charging and discharging the capacitance by adding and removing majority carriers to the edges of the space-charge layer, known as depletion or dielectric capacitance. $C_{eb\text{-}np}$ is given by (541.10) with appropriate extension and change of subscripts for use by a BJT:

$$C_{eb\text{-}pn} = (q^2/2kT)2n_iX_{EB}\exp(qV_{EB}/2kT)A_E$$

$$+ 2qn_i(dX_{EB}/dV_{EB})[\exp(qV_{EB}/2kT)-1]A_E. \qquad (741.17)$$

From (541.11), the dielectric or depletion capacitance is given by

$$C_{eb\text{-}d} = \varepsilon A_E/X_{EB}. \qquad (741.18)$$

The minority carrier charge storage capacitance in the thin quasi-neutral emitter and base layers are similar to those of the thick p/n junction diode, (541.8) and (541.9), but the thick diode solutions must be modified to take into account the thinness of the emitter and base layers or $X_E < L_E$ and $X_B << L_B$. They must be rederived from the d.c. solutions of the minority carrier (holes) concentration in the base (733.5) and a similar equation for minority carriers (electrons) in the emitter. Thus, using (733.5), the stored minority carrier charge densities (areal density in Coulomb/cm^2) for the p/n/p BJT are:

$$Q_B(V_{EB}) = \int_0^{X_B} q[P(x)-P_N]dx = \int_0^{X_B} qP_N[\exp(qV_{EB}/kT)-1]\frac{\sinh[(X_B-x)/L_p]}{\sinh[(X_B/L_p)]}\,dx$$

$$= q(P_B/L_B)[\exp(qV_{EB}/kT) - 1]\,\frac{\cosh(X_B/L_B) - 1}{\sinh(X_B/L_B)}$$

$$= q(P_B/L_B)[\exp(qV_{EB}/kT) - 1]\,\tanh(X_B/2L_B) \qquad (741.19)$$

$$\approx q(P_BX_B/2)[\exp(qV_{EB}/kT) - 1] \quad \text{if } X_B<<L_B. \qquad (741.19A)$$

The subscript B is used to denote the parameters associated with the minority carrier in the base layer. Thus, the minority-carrier (holes in p/n/p) diffusion length in the n-type quasi-neutral base layer is denoted by $L_B = \sqrt{D_B\tau_B}$ which in this case is $L_B \equiv L_p = \sqrt{D_p\tau_p}$ and the minority carrier (holes) concentration in the quasi-neutral base is denoted by P_B which in this case is $P_B \equiv P_N$. The expression for the minority carriers (electrons) stored in the p-type quasi-neutral emitter layer is obtained by replacing B by E and P by N in (741.19) and (741.19A),

$$\begin{aligned} Q_E(V_{EB}) &= q(N_E/L_E)[\exp(qV_{EB}/kT) - 1]\,\frac{\cosh(X_E/L_E) - 1}{\sinh(X_B/L_B)} \\ &= q(N_E/L_E)[\exp(qV_{EB}/kT) - 1]\ \tanh(X_E/2L_E) && (741.20) \\ &\approx q(N_E X_E/2)[\exp(qV_{EB}/kT) - 1] \quad \text{if } X_E \ll L_B. && (741.20A) \end{aligned}$$

The approximate results given by (741.19A) and (741.20A) can also be obtained directly by noting that the minority carrier density varies linearly with position when recombination loss is very small [Fig.733.1(d)] and has the equilibrium value at the emitter contact and at the base-edge of the base-collector junction. Thus, the stored charge is the area under the right triangle given by

$$Q_B = (1/2)q[P(0)-P_B]X_B = (1/2)qP_B[\exp(qV_{EB}/kT)-1]X_B \qquad (741.21A)$$

and

$$Q_E = (1/2)q[N(0)-N_E]X_E = (1/2)qN_E[\exp(qV_{EB}/kT)-1]X_E. \qquad (741.21B)$$

These are in agreement with the thin-layer approximation given by (741.19A) and (741.20A) and are frequently used by other authors as the starting point in the charge-control analysis. However, the exact small-signal solutions at all frequencies can be written down from (741.19) and (741.20) using the rule of (740.8C) of replacing τ_B by $\tau_B/(1+j\omega\tau_B)$, and τ_E by $\tau_E/(1+j\omega\tau_E)$. Note that the thin layer approximation can no longer be made because the electrical thickness of the emitter and base layer appears thick at high frequencies.

The thin emitter and base approximations at low frequencies ($X_E \ll L_E$ and $X_B \ll L_B$) are usually satisfied in high performance BJTs. The thin emitter condition is also known as the transparent emitter which is necessary to obtain high power-conversion efficiency in solar cells and high photon detection efficiency in photovoltaic detectors.

The charge storage capacitance in the quasi-neutral base is then

$$\begin{aligned} C_b &= A_E \partial Q_B/\partial V_{EB}\Big|_{V_{CB}} && (741.22) \\ &= A_E q(P_B/L_B)[(q/kT)\exp(qV_{EB}/kT)]\tanh(X_B/2L_B) && (741.23) \\ &\approx A_E(q^2P_BX_B/2kT)\exp(qV_{EB}/kT) \quad \text{if } X_B \ll L_B, && (741.23A) \\ &= g_b(X_B{}^2/2D_B) \equiv g_b t_B && (741.23B) \end{aligned}$$

where

$$t_B = X_B{}^2/2D_B = C_b/g_b \equiv 1/\omega_B. \qquad (741.23C)$$

t_B is the small-signal time constant of charging and discharging the quasi-neutral base layer by the minority carriers via diffusion into and out of the layer. ω_B is the corresponding angular frequency to be called the **base-charge cut-off frequency**. Note that many authors have used the MIT-SEEC notation, $\tau_F = X_B^2/2D_B$, instead of t_B. The τ_F notation causes confusion with the minority carrier recombination lifetime in the base, τ_B. We shall adhere to t_B which also conforms with the IEEE symbol convention of using Greek characters for fundamental material properties, such as μ_B, μ_n, τ_B and τ_n; and English characters for device, circuit-derived, and combined parameters, such as D_B, D_n, L_B, L_n, L_{Debye}, t_1, t_B, t_E, t_{EB}, and t_{CB}.

Similarly, the charge storage capacitance in the quasi-neutral emitter is

$$C_e = A_E \partial Q_E/\partial V_{EB}\Big|_{V_{CB}} \tag{741.24}$$

$$= A_E q(N_E/L_E)[(q/kT)\exp(qV_{EB}/kT)]\tanh(X_E/2L_E) \tag{741.25}$$

$$\simeq A_E(q^2 N_E X_E/2kT)\exp(qV_{EB}/kT) \quad \text{if } X_E \ll L_E, \tag{741.25A}$$

$$= g_e(X_E^2/2D_E) \equiv g_e t_E \tag{741.25B}$$

where

$$t_E = X_E^2/2D_E \qquad = C_e/g_e \equiv 1/\omega_E. \tag{741.25C}$$

t_E is the small-signal time constant of charging and discharging the quasi-neutral emitter layer by minority carrier diffusion into and out of the layer. ω_E is called the **emitter-charge cut-off frequency**.

Normally, $C_b = g_b t_B \gg C_e = g_e t_E$. This comes from $g_b \gg g_e$ because the emitter dopant impurity or majority carrier concentration is much higher than the base to give a high injection efficiency and beta. This offsets $t_E > t_B$ due to lower minority carrier diffusivity in the emitter owing to the higher emitter dopant concentration. For example, consider an Si n/p/n BJT with $X_E = X_B = 10^{-4}$cm, $D_E = D_p = 1\text{cm}^2/\text{s}$, $D_B = D_n = 10\text{cm}^2/\text{s}$, $P_E = n_i^2/N_{DD} = 10^{20}/10^{19} = 10\text{cm}^{-3}$, and $N_B = n_i^2/N_{AA} = 10^{20}/10^{17} = 10^3\text{cm}^{-3}$. Then, we get $t_E = 0.5\times10^{-8}\text{s} \gg t_B = 0.5\times10^{-9}\text{s}$, but $g_e/g_b = (D_E N_B X_B) \div (D_B P_B X_B) = (1\times10\times1) \div (10\times10^3\times1) = 10^{-3}$ or $g_b \gg g_e$. Thus, $C_e/C_b = 10^{-2}$ or $C_b \gg C_e$.

A similar procedure can be used to compute the capacitance elements of the collector-base junction, C_{cb}. The full algebra just presented needs not be carried out since the collector junction is reverse biased in small-signal amplifier and oscillator applications and the only dominant capacitance term is the dielectric capacitance of the depleted collector-base junction space-charge layer given by

$$C_{cbt} = A_C \varepsilon/X_{CB} \tag{741.26}$$

where

$$X_{CB} = \sqrt{2\varepsilon(V_{CB-bi} - V_{CB})/qN_{HC}} \tag{741.27A}$$

and

$$N_{HC} = N_{CC}N_{BB}/(N_{CC}+N_{BB}). \tag{741.27B}$$

The capacitance current flowing into the collector terminal due to the decrease of the base stored charge from the Early effect is

$$\begin{aligned} C_a &= \partial Q_B/\partial V_{CB} = A_E(qP_B/2)(X_B/V_{CB})(\partial \log_e X_B/\partial \log_e V_{CB})\cdot \exp(qV_{EB}/kT) \\ &= A_E(qP_B/2)(-X_B/V_{CB})m\cdot \exp(qV_{EB}/kT) \\ &= |-mkT/qV_{CB}|C_b \ll C_b. \end{aligned} \qquad (741.28)$$

Collecting all the terms, the total capacitance across the emitter-base and collector-base junctions shown in Fig.741.1(d) are respectively

$$C_{eb} = C_e + C_{ebt} + C_b \qquad (741.29)$$

and

$$C_{cb} = C_{cbt} + |mkT/qV_{CB}|C_b. \qquad (741.30)$$

After adding these capacitances, the CBss-hTee equivalent circuit becomes implicit as indicated in Fig.741.1(d), namely, the dependent current source, $\alpha_{fo}i_{e'}$, now depends on an internal and inaccessible or not directly measurable current, $i_{e'}$, which passes through the conductance g_b; so named implicit. In order to transform the current source to become explicit, i.e. being dependent on an accessible terminal current that can be measured by an experimenter, the following simple circuit analysis is undertaken.

$$i_{e'} = i_e g_{eb}/(g_{eb}+j\omega C_{eb}) = i_e/(1 + j\omega/\omega_\alpha) \qquad (741.31)$$

then

$$\alpha_{fo}i_{e'} = \alpha_{fo}i_e/(1 + j\omega/\omega_\alpha) \equiv \alpha_f i_e \qquad (741.32)$$

where

$$\alpha_f \equiv \alpha_{fo}/(1 + j\omega/\omega_{\alpha'}) \equiv \alpha_B/(\gamma_{eo}^{-1} + j\omega/\omega_\alpha) \qquad (741.32A)$$

$$2\pi f_{\alpha'} = \omega_\alpha' \equiv g_{eb}/C_{eb} \qquad \textbf{True -3db alpha cut-off frequency}$$

$$= (g_e+g_{ebt}+g_b)/(C_e+C_{ebt}+C_b) \qquad (741.33A)$$

$$= \gamma_{eo}/[t_B + (C_e+C_{ebt})/g_b] \equiv \gamma_{eo}\omega_\alpha = 2\pi\gamma_{eo}f_\alpha \qquad (741.33A1)$$

$$2\pi f_\alpha = \omega_\alpha \equiv 1/[t_B + (C_e+C_{ebt})/g_b] \qquad \textbf{"Alpha cut-off frequency"} \quad (741.33B)$$

$$= \omega_{\alpha'}/\gamma_{e0} = 2\pi f_{\alpha'}/\gamma_{eo} \qquad (741.33B1)$$

and

$$2\pi f_B = \omega_B \equiv 1/t_B \equiv g_b/C_b$$

$$= 2D_B/X_B^2. \qquad \textbf{Base-charge cut-off frequency} \quad (741.33C)$$

Several **characteristic** frequencies are introduced in the preceding derivation of the CBss-Tee and CBss-hTee models of an **intrinsic** BJT. They are defined in order to distinguish their physical origins and to provide a precise comparison with the exact solution. Such a comparison is not readily available in existing textbooks.

In the past and latest research, engineering and trade reports and in recent journal articles, there are frequent inconsistencies on the usage of the various characteristic frequencies to gauge and benchmark the lastest technology developments of submicron 100GHz/10ps Si, GaAs and heterojunction BJTs. These frequencies are now defined to indicate their practical and theoretical significance.

$\omega_{\alpha'}$ or $f_{\alpha'}=\omega_{\alpha'}/2\pi$ defined in (741.33A) is the frequency at which the magnitude of α_f, **CBss h_{21}**, or the **CBss short-circuit-output current gain** or current transfer ratio, defined by $i_c/i_e = I_c(\omega=\omega_{\alpha'})/I_e$ at $v_{cb}=0$, drops 3-db from the low frequency value, $\alpha_{f0}=\alpha_B\gamma_{e0}$, to $\alpha_{f0}/\surd 2 = 0.707\alpha_{f0}$. It is known as the (true) **alpha (minus 3-db) cut-off frequency**. It can be measured experimentally but seldom is in characterizing the latest developmental 10-100GHz BJTs because the frequency dependence of the CEss short-circuit-output current gain or beta can be measured at a much lower frequency, 100-times lower $\approx \omega_{\alpha'}/\beta_F \approx \omega_{\alpha'}/100$, to give a computed $\omega_{\alpha'}$. Lower frequency measurements are much easier to implement and give much better accuracy.

ω_α or $f_\alpha=\omega_\alpha/2\pi$ defined in (741.33B) and given by $\omega_\alpha=\omega_{\alpha'}/\gamma_{e0}$ is the often called **alpha cut-off frequency** found in the literature and textbooks. It is nearly equal to the true -3db alpha cut-off frequency $\omega_{\alpha'}$ in most BJTs with high β_F because they have nearly unity emitter injection ratio, $\gamma_{e0}=1$.

ω_B $(=1/t_B=2D_B/X_B^2)$ or f_B defined in (741.33C) and (741.23C) is the charge-control approximation of the minus 3db frequency of the **small-signal base transport factor** of the minority-carrier (holes) diffusion and drift signal through the quasi-neutral (n-type) base layer. It is known as the **base-charge cut-off frequency**. t_B is precisely also the RC time constant of the base diffusion conductance and capacitance, g_b and C_b in parallel, as indicated by (741.23B). ω_B or t_B is an internal parameter of the BJT and cannot be measured directly from the transistor terminals. It is defined by

$$\alpha_b(\omega=\omega_B) = \alpha_b(\omega=0)/\surd 2 \equiv \alpha_B/\surd 2 \tag{741.34}$$

where small-signal low-frequency or zero-frequency base transport factor, $\alpha_b(\omega=0)$, is identical to the d.c. base transport factor defined in (733.21D) and expanded by Taylor series in (737.2B):

$$\alpha_B = \operatorname{sech}(X_B/\surd D_B\tau_B)$$

$$\approx 1 - X_B^2/2D_B\tau_B \equiv 1 - t_B/\tau_B \equiv 1 - \omega_B t_B \tag{741.35}$$

where

$$t_B = 1/\omega_B = X_B^2/2D_B. \tag{741.35A}$$

Similarly, $\omega_E=1/t_E=2D_E/X_E^2$ or $f_E=\omega_E/2\pi$, defined in (741.25B) and (741.25C), is the charge control approximation of the -3db frequency of the small-

signal emitter transport factor of the minority-carrier (electrons) diffusion and drift signal through the quasi-neutral (p-type) emitter layer. It is known as the **emitter-base cut-off frequency**. t_E is precisely also the RC time constant of the parallel emitter diffusion conductance and capacitance, g_e and C_e, as indicated by (741.25B). ω_E or t_E is also an internal parameter of the BJT and cannot be measured directly from the transistor terminals.

These charge-control results are excellent approximations to the exact results from the small-signal expansion of the electron and hole diffusion equations or using the substitution procedure of replacing τ_B and τ_E in the d.c. equations by $\tau_B/(1+j\omega\tau_B)$ and $\tau_E/(1+j\omega\tau_E)$ described in (740.8C). This gives the square of the normalized complex base and emitter quasi-neutral layer thicknesses which are

$$Z_B{}^2 = [X_B{}^2(1+j\omega\tau_B)/D_B\tau_B] = 2[(t_B/\tau_B) + j\omega t_B] \qquad (741.36A)$$

and

$$Z_E{}^2 = [X_E{}^2(1+j\omega\tau_E)/D_E\tau_E] = 2[(t_E/\tau_E) + j\omega t_E]. \qquad (741.36A)$$

The exact and 1-pole approximation of the two-port parameters are then:

$$y_{mf} = y_{21} = g_{mfo}(Z_B/\sinh Z_B) \qquad (741.37)$$

$$\approx g_{mfo}\exp(j\theta_{gm})/(1 + j\omega/\omega_{gm}) \approx g_{mfo} = \alpha_B g_b \qquad (741.37A)$$

where

$$\omega_{gm} = 4.5903/t_B = 4.590(2D_B/X_B{}^2), \qquad \text{nev result?} \qquad (741.37B)$$

and

$$\theta_{gm} \le 0.1218\omega_{gm}t_B \quad \text{when } \omega \le \omega_{gm}; \qquad \text{nev result?} \qquad (741.37C)$$

and

$$y_{ee} = y_{11} = j\omega C_{ebt} + g_e Z_E/\tanh Z_E + g_{ebt} + g_b Z_B/\tanh Z_B \qquad (741.38)$$

$$\alpha_b = \mathrm{sech}(Z_B) \qquad (741.39)$$

$$\approx \alpha_{bo}\exp(j\theta_{\alpha b})/(1 + j\omega/\omega_{\alpha b}) \qquad (741.39A)$$

where

$$\alpha_{bo} \equiv \alpha_B = \mathrm{sech}(X_B/\sqrt{D_B\tau_B}) \approx 1-X_B{}^2/2D_B\tau_B \equiv 1-t_B/\tau_B \qquad (741.39B)$$

$$\omega_{\alpha b} = 1.2162/t_B = 1.2162(2D_B/X_B{}^2) \qquad \text{Middlebrook} \qquad (741.39C)$$

and

$$\theta_{\alpha b} < 0.1736\omega_{\alpha b}t_B \quad \text{when } \omega < \omega_{\alpha b}. \qquad \text{nev result?} \qquad (741.39D)$$

Middlebrook derived the first approximation formulae of the base alpha [R.D.Middlebrook, **An Introduction to Junction Transistor Theory**, John Wiley & Sons, 1957, chapter 11, pp.192-209] which gave accurate frequency dependences of both the phase and amplitude. The **Middlebrook alpha (Middlebrook frequency**, $\omega_{\alpha b}$) is

$$\alpha_b = \alpha_{bo}[1 - j(0.214)(\omega/\omega_{\alpha b})]/[1 + j(1.04)(\omega/\omega_{\alpha b})]. \qquad (741.39E)$$

The CB current transfer ratio or alpha is

$$\alpha_f = -\ h_{21-cb} = -\ y_{21}/y_{11}$$

$$= 1/\{\cosh Z_B + [(g_e Z_B/\tanh Z_B) + g_{ebt} + j\omega C_{te}]/y_{21}\} \qquad (741.40)$$

$$\approx \alpha_B/\{1+j(\omega/\omega_{\alpha b}) + [(g_e+g_{ebt}+j\omega C_{te})/g_b][1+j(\omega/\omega_{gm})]\} \qquad (741.40A)$$

$$\approx \alpha_B/\{\gamma_{eo}{}^{-1} + j(\omega/\omega_\alpha)\} \qquad (741.40B)$$

where

$$1/\omega_\alpha = t_B/1.2162 + (C_e+C_{te})/g_b. \qquad (741.40C)$$

The approximation to the exact small-signal complex alpha given by (741.40B) above and the charge-control complex alpha given by (741.32A) are identical in form except that the factor 1.2162 (or 2.4324/2) in the exact solution gives a 21.62% higher base-diffusion cutoff frequency, $\omega_{\alpha b}=2.4324D_B/X_B{}^2=1.2162\omega_B =1.2162/t_B$, than the charge-control value of $\omega_B=2D_B/X_B{}^2=1/t_B$.

In addition to the two contributions to the alpha cutoff frequency shown in (741.40C), there is a small-signal attenuation and drift delay, $t_{cb\text{-}sat} = X_{CB}/2\Theta_{sat}$, from minority carrier signal transmitting through the depleted collector-base junction space-charge layer at the saturation drift velocity, $\Theta_{sat} = v_d$(high-field) $\approx 10^7$cm/s shown in Fig.314.1. Thus, (741.40C) is modified and becomes

$$1/\omega_\alpha = t_B/1.2162 + (C_e+C_{ebt})/g_b + (X_{CB}/2\Theta_{sat}). \qquad (741.41)$$

CBss-Tee Equivalent Circuit Model of Real BJT

The model of a real BJT transistor must include the parasitics that are not included in the 1-d intrinsic transistor model just described. The parasitics include resistances against the majority carrier current flows in the thin quasi-neutral layers. These are the collector resistance due to the vertical Si bulk, $r_{c'}$; the lateral resistance of the thin quasi-neutral base layer, $r_{b'}$; and the series resistance of the quasi-neutral emitter layer, $r_{e'}$; which also include the series contact resistance of the conductor/Si interface at the emitter, base and collector terminals. The parasitics include also the shunt capacitances between the neighboring and overlapping interconnect conductors and the Si layers in the forms of MOS, MOM, and SOS parallel plate and edge-fringe capacitances; the inductance of the interconnection lines (in monolithic BJT for integrated circuits) and off-chip wiring; and the leakage current, shunt resistances and shunt capacitances of parasitic rectifying junctions, such as the underlap collector-base diode. To distinguish these parasitic circuit elements from the circuit elements of the intrinsic transistor, the parasitics are labeled by a lower case letter and lower case subscript, which is also primed, such as $r_{b'}$ indicated above. The intrinsic or internal nodes are also primed, such as b', while the external nodes are unprimed such as b. If the IEEE notation convention is strictly adhered to, double subscripts should be used for the parasitic elements to indicate their position between two nodes, such as $r_{b'b}$ and $r_{e'e}$, just like the intrinsic element such as g_{eb} and g_{cb}. However, single subscript with a prime has been used traditionally. Due to a limited number of symbols, r and g may not be reciprocals of the same element. In some cases, r is a series element while g is a different shunt element.

Following the partitioning methodology of treating a three-dimensional transistor by decomposition into paralleled one-dimensional transistor and diodes with distributed RC interconnecting lines, these one-lump parasitics are added to the terminals of the intrinsic transistor of Fig.741.1(d) to give the complete CBss-hTee equivalent circuit model of a real 3-d BJT shown in Fig.741.1(e). The distributed RC lines, parasitic inductances of interconnection lines and wires are not shown in this one-lump approximation. The size of the parasitics and their structural and material origins are described in the following paragraphs using Fig.741.2.

For a well-designed high-frequency transistor, $r_{e'}$ is small. The emitter layer is heavily doped and has high conductivity which keeps down the lateral and transverse resistances from the very thin layer (0.1μm) in the latest self-aligned BJTs with 10 to >100GHz cut-off frequencies. $r_{e'}$ is also reduced by using a high conductivity metal/metal-silicide M/M_xSi overlay (such as Al/silicide or some refractory metal/silicide) on the poly-Si/n+emitter contract layer of the [M/M_xSi/poly-Si/n+emitter]/p/n-/n+ BJT structure. The main source of $r_{e'}$ is the interfacial contact resistance between the layers, such as the contact M_xSi/poly-Si. The contact resistance is invariably increased by a residual thin interfacial silicon-dioxide which was not dissolved during the contact metallization reaction. Semiconductor metallurgy has advanced swiftly, giving $<20\Omega/\mu m^2$ emitter contact resistances.

The collector body series resistance $r_{c'}$ from the silicon substrate can be reduced using an extension of Early's LDC (Lowly Doped Collector) n+/p/i/n drift transistor structure. An extended n+/p/n-/n+/p- structure is shown in Fig.741.2 which has an additional n+ layer known as the **buried collector**. The /n-/ collector layer comes from a thin and lowly doped epitaxial Si layer grown on a low conductivity /p- substrate with a preimplanted or prediffused /n+/ buried layer. The thickness and impurity doping (or resistivity) of the /n-/ epitaxial layer are designed so that the final /n-/ layer is **punched through** by the collector-base junction space-charge layer at the d.c. operating collector-base junction voltage. The punch-through layer is shaded in Fig.741.2 and it leaves no undepleted low-field /n-/ layer which would have given a high $r_{c'}$. The resistivity and thickness of the /n-/ layer can also be selected to give a prescribed collector-base junction breakdown voltage.

Although $r_{e'}$ and $r_{c'}$ can easily be reduced and minimized by proper structure design and doping selection of the emitter and collector layers, the base resistance,

$r_{b'}$, cannot be eliminated and can only be minimized because it comes from the majority carrier drift current in the base layer due to recombination of the injected minority carriers with the existing majority carriers, and because this majority carrier current must flow through the thin base laterally in order to reach the base contact and goes out of the base terminal to the external circuit. The physical location of the lateral base resistance is illustrated in Fig.741.2. The base resistance, $r_{b'}$ = resistivity·length/area = $(q\mu_p N_{BB})^{-1}\cdot(\frac{1}{2}W_B+W_{EB})/(X_B Z_B)$, cannot be reduced without limit by increasing the thickness X_B and base dopant concentration N_{BB} because the base layer must be kept thin and the dopant concentration sufficiently low to have high minority carrier mobility and diffusivity in order to give high cut-off frequencies ($\omega_B=2D_B/X_B^2=1/t_B$, $\omega_{\alpha b}=2.43D_B/X_B^2$), and high d.c. gain or beta ($\beta_F=\tau_B/t_B=2D_B\tau_B/X_B^2$). Recent (1989-1991-onwards) world-wide development efforts, started at IBM in the mid-1980's by T.H.Ning, D.Tang, and others, have used the self-aligned, arsenic-doped poly-Si source for emitter diffusion and contact to reduce the emitter width W_B and emitter-base contact gap, W_{EB}. These efforts have produced Si BJTs with very low $r_{b'}$, nearly 100GHz cut-off frequency, and sub-ten picoseconds switching delay.

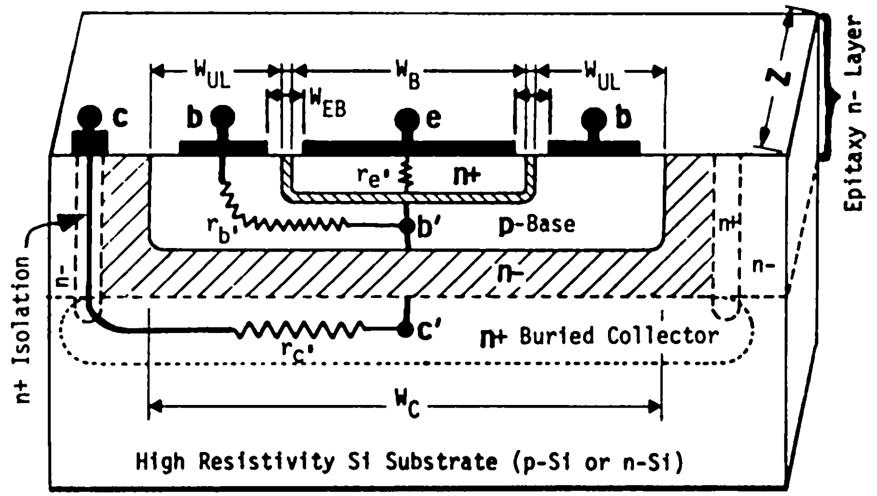


Fig.741.2 Cross-section of real BJT showing lateral base resistance and series collector resistance.

The underlap collector-base junction diode, indicated by the width W_{UL} in Fig.741.2, must also be included. Since for small-signal applications the collector junction is always reverse biased, the underlap diode contributes only an additional collector-base capacitor in the small-signal equivalent circuit, shown as C_{ul} in Fig.741.1(e). It can be reduced by narrowing the underlap width W_{UL} and lowering the dopant concentration in the underlap region. Self-aligned emitter technology has reduced W_{UL} very significantly.

742 Maximum Frequency of Oscillation of BJT

Fabrication technology of state-of-the-art ultrahigh frequency BJTs has been compared and gauged by several figure-of-merits that can be readily measured experimentally. These are: the -3db cut-off frequency of beta or CE output short-circuit current gain, the frequency of unity beta (also known as bandwidth of the gain-bandwidth product), the frequency of unity maximum-available-power-gain, and **the maximum frequency of oscillation**, f_{Max}. There are also several figure-of-merits for large-signal switching applications which are discussed in sections 75n.

The Gibbons Frequency

The maximum-frequency-of-oscillation f_{Max} will be derived in this section using the CBss equivalent circuit. f_{Max} has been used almost since the beginning of bipolar transistor development in the early 1950's. Only four parts are needed to measure f_{Max}: a high-Q or low-loss (at $f \propto f_{Max}$) inductor L, a high-Q or low-loss capacitor C (both from a microwave waveguide or strip line), a d.c. power supply, and a millimeter wave detector. Furthermore, because f_{Max} is a function of both ω_α and $r_{b'}C_c$, it also gives an experimental $r_{b'}$ which helps to check the design optimization of the BJT via reduction of $r_{b'}$ and C_c.

The formulae of f_{Max} can be readily derived from the CBss-hTee model given in Fig.741.1(e). Since the highest frequency is the objective, all the resistances and parasitics are deleted except r_b' and C_c. Then, Fig.741.1(e) simplifies to Fig.742.1 in which an external variable inductor L and variable capacitor C are added which are adjusted to give the maximum value of the frequency of oscillation. If C_c is very high, a second inductor between c and e is needed to tune out C_c.

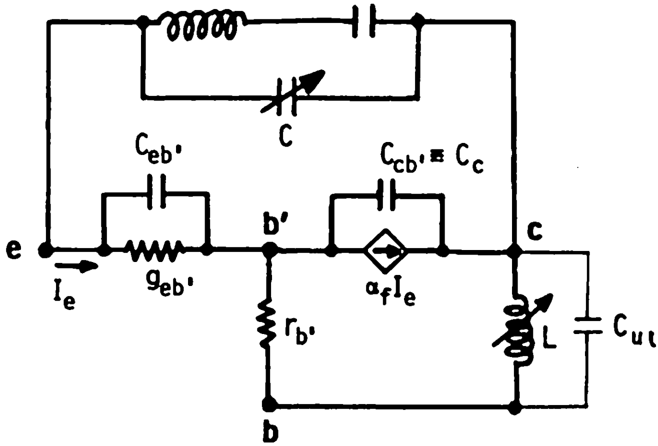


Fig.742.1 The equivalent circuit for measuring the maximum frequency of oscillation of a BJT.

The maximum frequency of oscillation is the condition of infinite gain. Hence, it is independent of the equivalent circuit configuration because a finite voltage is developed at the internal nodes and terminals of the BJT without any input applied to any terminals, i.e. infinite gain. It corresponds to zero input conductance between b' and c, $G_{b'c}=0$, if the emitter-base junction resistance, $g_{eb'}^{-1}$, is small. If not, then $1/G_{b'c}$ is just negative enough to balance out $\mathrm{Re}[(g_{eb}+j\omega C_{eb})^{-1}]$. The negative input capacitance (negative due to L between c and b) is tuned out by adjusting C. Thus, the total admittance (transistor plus those from C and L) between b' and c is zero. Then, a small current noise or disturbance, δI_e, will produce an infinite $V_{b'c}$ and the circuit will oscillate. To ease the derivation of f_{Max}, a high current is assumed such that the emitter-base junction loss, $g_{eb'}$, can be neglected (shorted or $g_{eb'}=\infty$) as a first approximation. The underlap capacitance, C_{ul}, and other parasitics can be tuned out by L and C; hence, they are dropped from the equivalent circuit. Thus, we only need to consider the admittance between terminals b' and c to obtain the f_{Max}. The idea of -90° phase when $G_{b'c}\leq 0$ was first used by James F. Gibbons (Stanford 1962, current Stanford engineering dean) whose derivation of f_{Max} followed a more tedious but graphic, phasor route. Setting $G_{b'c}=0$ tremendously simplifies the algebra of a rigorous derivation of f_{Max} which we shall now give. To get a negative real part from $Y_{b'c}$ or to make its phase angle greater than |-90°|, the external inductor L between the collector-base terminals is varied in order to tune out C_c and the negative phase shift from the current gain, $\alpha_f=\alpha_{fo}/(1+j\omega/\omega_{\alpha'})$. The node equation at b' is

$$I_e = \alpha_f I_e + V_{b'c}[j\omega C_c + 1/(r_{b'}+j\omega L)] \qquad (742.1)$$

giving

$$Y_{b'c} = I_e/V_{b'c}$$

$$= [j\omega C_c + 1/(r_{b'}+j\omega L)]/(1-\alpha_f) \qquad (742.2)$$

$$= [j\omega C_c + 1/(r_{b'}+j\omega L)](1+j\omega/\omega_{\alpha'})/(1-\alpha_{fo}+j\omega/\omega_{\alpha'}). \qquad (742.3)$$

Since the transistor has a near unity alpha or large beta at low frequencies, $1-\alpha_{fo} << \omega_{Max}/\omega_{\alpha'}$, thus, $\omega_{Max} >> \omega_{\alpha'}/(1-\alpha_{fo}) = \omega_{\alpha'}/\beta_{fo} \equiv \omega_\beta$, then

$$Y_{b'c} \cong [j\omega C_c + 1/(r_{b'}+j\omega L)](-j\omega_{\alpha'}/\omega + 1) \qquad (742.3A)$$

$$= \omega_{\alpha'}C_c + j\omega C_c + (1-j\omega_{\alpha'}/\omega)/(r_{b'}+j\omega L) \qquad (742.4)$$

$$= \omega_{\alpha'}C_c + j\omega C_c + \frac{r_{b'}-\omega_{\alpha'}L - j(\omega L+\omega_{\alpha'}r_{b'}/\omega)}{r_{b'}^2 + \omega^2L^2}. \qquad (742.5)$$

The condition of oscillation is when the real part of $Y_{b'c}$ is zero,

$$\mathrm{Re}[Y_{b'c}] = \omega_{\alpha'}C_c + (r_{b'}-\omega_{\alpha'}L)/(r_{b'}^2+\omega^2L^2) = 0 \qquad (742.6)$$

or

$$0 = \omega_{\alpha'}C_c(r_{b'}^2 + \omega^2L^2) + (r_{b'} - \omega_{\alpha'}L). \qquad (742.7)$$

To seek the maximum frequency of oscillation by adjusting L, we set $\partial\omega/\partial L=0$ in (742.7) and obtain **Gibbons' formula** of f_{Max}:

$$\omega = \omega_{Max} = \left[\frac{\omega_{\alpha'}}{4r_{b'}C_c(1+\omega_{\alpha'}r_{b'}C_c)}\right]^{1/2} \qquad \textbf{(Gibbons' formula)} \qquad (742.8)$$

$$= \left[\frac{\omega_{t-}}{4r_{b'}C_c}\right]^{1/2} \qquad (742.8A)$$

where ω_{t-} will be known as the **Gibbons frequency** and is given by

$$\omega_{t-} = t_{t-1} = 2\pi f_{t-} \qquad \textbf{(Gibbons' frequency)}$$

$$= [\omega_{\alpha'}{}^{-1} + r_{b'}C_c]^{-1} \qquad (742.8B)$$

$$= \gamma_{eo}/[t_B/1.2162 + (C_e+C_{ebt})/g_b + X_{CB}/2\Theta_{sat} + C_c r_{b'}]^{-1} \qquad (742.8C)$$

$$\neq \omega_t = \gamma_{eo}/[t_b/1.2162 + (C_e+C_{ebt})/g_b + X_{CB}/2\Theta_{sat} + C_c/g_{eb}]^{-1} \qquad (744.7)$$

and

$$f_{Max} = \left[\frac{f_{t-}}{8\pi r_{b'}C_c}\right]^{1/2} = \frac{1}{4\pi}\left[\frac{1}{r_{b'}C_c t_{t-}}\right]^{1/2}. \qquad (742.9)$$

Note, the Gibbons frequency, f_{t-}, defined in (742.8C) is not equal to the commonly used unity beta frequency, f_t, or bandwidth, $f_{bw}(=f_t)$, (744.7) listed above, which is the CE current-gain bandwidth to be derived in section 744 using the CE equivalent circuit model. The difference lies in the term $r_{b'}C_c$ of ω_{t-} which is replaced by C_c/g_{eb} in ω_t. This difference is very significant in state-of-the-art GHz BJT whose $r_{b'}C_c$ is comparable to $\omega_{\alpha'}{}^{-1}$. It is incorrect to use f_{bw} (or f_t) in place of f_{t-} to compute f_{Max} as done by recent BJT developers at various laboratories. Table 755.1 will show the tremendous difference between $f_t(=f_{bw}=26.4\text{GHz})$ and $f_{t-}(=4.42$ to $22.1\text{GHz})$ in self-aligned submicron Si BJTs when $r_{b'}C_c$ is reduced.

As a numerical example to illustrate order of magnitudes, we consider the 1961 Si n/p/n 2N708 which was the first high-frequency Si transistor ever produced. The data sheet of 2N708 gives $f_\alpha = 300\text{MHz} \propto f_t$, $C_c=6.0\text{pF}$ and $r_{b'}=50\Omega$, giving $r_{b'}C_c=300\text{ps}$. Neglecting the emitter delay by assuming a high emitter or collector current so that $(C_e+C_{ebt})/g_b \le 10\text{ps}$, and neglecting collector-base junction transit-time delay, $X_{CB}/2\Theta_{sat} \propto 1\text{ps}$, then, $t_t = (2\pi\cdot 300\text{x}10^6)^{-1}\text{x}10^{112} + 300\text{ps} = 530\text{ps} + 300\text{ps} = 830\text{ps}$ and f_{Max} is

$$f_{Max} = \sqrt{300\text{x}10^6/[8\pi\text{x}50\text{x}6\text{x}10^{-12}(1+2\pi\text{x}300\text{x}10^6\text{x}50\text{x}6\text{x}10^{-12})]}$$

$$= \sqrt{300\text{x}10^6\text{x}1.32\text{x}10^8/(1+0.5655)} = 159\text{ MHz}.$$

A better collector design, such as using a high resistivity epitaxial layer shown in Fig.741.2, could reduce the collector capacitance by 10 times to 0.6pF and increase the maximum frequency of oscillation three times to 477 MHz.

The results just obtained give two formulae to compute the transistor parameters from experimental data. Let's assume that f_{Max} and L are both measured during the experiment. Then

$$C_c = 1/(2\omega_{Max}{}^2 L) \qquad (742.10)$$

and

$$\omega_{\alpha'} C_c r_{b'}{}^2 + r_{b'} = \omega_{\alpha'} L/2. \qquad (742.11)$$

$\omega_{\alpha'}$ can be computed from measurements of $\beta_f(\equiv I_c/I_b \equiv h_{21})$ versus ω whose -3db point gives $\omega_\beta \equiv (\omega_{\alpha'}{}^{-1} + C_c/g_{eb})^{-1}/\beta_{fo}$ if C_c/g_{eb} can be neglected (derived in section 744). Then, $r_{b'}$ can be computed from (742.11).

The formula of f_{Max} just derived was obtained by assuming negligible emitter-base admittance loss, i.e. $\mathrm{Re}[1/Y_{eb}]=0$. In practice, this may not be the case. However, the algebra involved to obtain an analytical f_{Max} is laborious when $r_{eb}(\equiv 1/g_{eb})$ is not neglected or not set equal to zero, $r_{eb}=0$. But, it is evident from simple physics that the f_{Max} formula just obtained in (742.9) with $r_{eb}=0$ is the highest possible f_{Max}. Recombination loss in the quasi-neutral emitter and base (g_e and g_b) and in the emitter-base junction space-charge layer (g_{ebt}) will reduce f_{Max}. If $r_{eb} \neq 0$, the reduction of f_{Max} due to $C_{eb} r_{eb}$ loss can be minimized or made to approach zero by adjusting L such that C is very small. When $C << C_{eb}$, one can conclude by an inspection of Fig.742.1 that f_{Max} will approach Gibbons' formula. In practice the explicit f_{Max} including r_{eb} is not derived but the highest f_{Max} is computed from measured f_t by varying $C_c(V_{CB})$, $C_{eb}(I_E)$, and $g_{eb}(I_E)$ via varying V_{CB} and I_E or $I_C(\equiv \alpha_F I_E \simeq I_E)$.

743 Common-Emitter Small-Signal Hybrid-Pi (CEss-Hπ) Model of BJT

In small-signal applications, BJTs are usually connected in the common-emitter configuration using the base terminal as the input in order to give high amplification of both signal current and voltage. The **CEss Hybrid-Pi** equivalent circuit model is the most frequently used model. It can be obtained from the CBss-Tee model using straightforward algebra or a two-port matrix transformation. It can also be obtained from the CE explicit d.c. hybrid-π model given in Fig.735.2(b) by taking the differential of the d.c. circuit elements and then adding the charge storage capacitances. These are left as exercise problems. Instead, we shall follow the charge-control procedure used to derive the CBss-Tee model in section 741 in which the conductance elements were first obtained by taking the differential of the original d.c. equations given by (733.19) for J_E, (733.20) for J_C, and their sum $J_B=-(J_E+J_C)$ for the base terminal. This gives the low-frequency small-signal equivalent circuit model. Then, the charge-storage capacitances, the parasitic series resistances and the shunt capacitances are added between appropriate nodes to give the complete high-frequency small-signal equivalent circuit model. This route of derivation helps to trace the physical origin of the circuit elements directly without having to go through an intermediate equivalent circuit such as the CBss-Tee if the CEss-Hπ were obtained from CBss-Tee by a 2-port network matrix transformation.

The a.c./d.c.-bias circuit, intrinsic low-frequency circuit, intrinsic high-frequency equivalent circuit, and the complete extrinsic high-frequency equivalent circuit are shown in Figs.743.1(a)-(d). They derived in the following subsections.

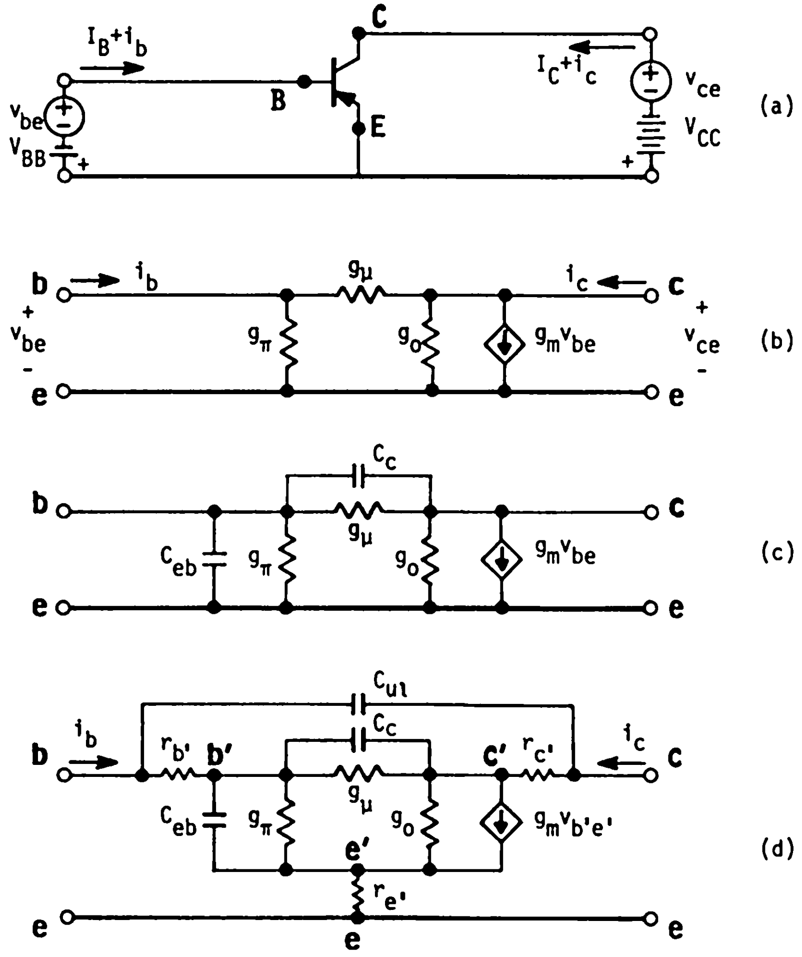

I_B+i_b
C
B
E
I_C+i_c
v_{be}
V_{BB}
+
v_{ce}
V_{CC}
+
(a)
b
i_b
g_μ
i_c
c
+
v_{be}
-
g_π
g_o
$g_m v_{be}$
+
v_{ce}
-
e
e
(b)
C_c
b
c
g_μ
C_{eb}
g_π
g_o
$g_m v_{be}$
e
e
(c)
C_{u1}
C_c
i_b
b
$r_{b'}$
b'
c'
$r_{c'}$
i_c
c
g_μ
C_{eb}
g_π
g_o
$g_m v_{b'e'}$
e'
$r_{e'}$
e
e
e
(d)

Fig.743.1 The common emitter BJT circuits. (a) The d.c. biases and the small-signal voltage sources. (b) The intrinsic CEss hybrid-pi low-frequency model. (c) The intrinsic CEss-Hπ high-frequency model with charge-control capacitances. (d) The complete CEss-Hπ model.

The Conductance Elements of the CEss-Hπ BJT Model

The d.c. bias and the small-signal voltage sources are shown in Fig.743.1(a). The low-frequency CEss hybrid-pi equivalent circuit, containing only the self conductance and transconductance elements, is shown in Fig.743.1(b). These conductances can be more readily derived by taking the differential of the original Shockley or SNS transistor equations of the emitter, base and collector d.c. currents rather than the extended Ebers-Moll d.c. equations. The original equations were given by (733.19) and (733.20) which are repeated for convenience:

$$J_E = (j_E+j_{EB}+j_B)[\exp(qV_{EB}/kT)-1] - \alpha_B j_B[\exp(qV_{CB}/kT)-1] \qquad (743.1)$$

$$J_C = (j_C+j_{CB}+j_B)[\exp(qV_{CB}/kT)-1] - \alpha_B j_B[\exp(qV_{EB}/kT)-1]. \qquad (743.2)$$

The base current is

$$\begin{aligned} J_B &= -(J_E + J_C) \\ &= -[j_E+j_{EB}+j_B(1-\alpha_B)][\exp(qV_{EB}/kT)-1] \\ &\quad + [j_C+j_{CB}+j_B(1-\alpha_B)][\exp(qV_{CB}/kT)-1]. \end{aligned} \qquad (743.3)$$

For small-signal operation as an amplifier with high gain, these equations can be further simplified by neglecting the leakage currents. Then, for a device area of A_E, we have

$$I_B = J_B A_E \approx -[j_E+j_{EB}+j_B(1-\alpha_B)][\exp(-qV_{BE}/kT)]A_E \qquad (743.4)$$

$$I_C = J_C A_E \approx \qquad -\alpha_B j_B[\exp(-qV_{BE}/kT)]A_E \qquad (743.5)$$

whose small-signal expansion gives

$$i_b = (\partial I_B/\partial V_{BE})\Big|_{V_{CE}} v_{be} + (\partial I_B/\partial V_{CE})\Big|_{V_{BE}} v_{ce} \qquad (743.6)$$

$$i_c = (\partial I_C/\partial V_{CE})\Big|_{V_{BE}} v_{ce} + (\partial I_C/\partial V_{BE})\Big|_{V_{CE}} v_{be}. \qquad (743.7)$$

These are the short-circuit admittance two-port equations. From their input and transfer conductances, the circuit elements of the hybrid-pi circuit elements of Fig.743.1(b) can be derived. Thus, by inspection (shorting e-c nodes and applying a test voltage to b-e nodes and measure i_c)

$$g_m - g_\mu = (\partial I_C/\partial V_{BE})\Big|_{V_{CE}} = q(-I_C/kT) \qquad (743.8)$$

where g_m is the low-frequency transconductance and g_μ is the feedthrough conductance. The other three resistance or conductance elements are computed as follows. The short-circuit input conductance is

$$g_\pi + g_\mu = (\partial I_B/\partial V_{BE})\Big|_{V_{CE}} = I_B(\partial \log_e I_B/\partial V_{BE})\Big|_{V_{CE}} \tag{743.9}$$

$$= - I_B\{(q/kT) - (\partial/\partial V_{BE})\log_e[j_{RE}+j_E+j_B(1-\alpha_B)]\Big|_{V_{CE}}\} \tag{743.10}$$

$$= [(g_m - g_\mu)/\beta_F]\{1 - (\partial/\partial V_{BE})\log_e[j_{RE}+j_E+j_B(1-\alpha_B)]\Big|_{V_{CE}}\} \tag{743.11}$$

$$= [(g_m - g_\mu)/\beta_F][1 + E_e] \tag{743.12}$$

where g_π is the input conductance. The not-evaluated derivative term, E_e, comes from the base-layer thickness modulation by the time-varying signal applied to the base-emitter junction (the a.c. emitter Early effect). E_e is generally negligible but it may become important in very thin base, very high frequency and very high speed transistors.

The short-circuit output conductance is given by

$$g_o + g_\mu = (\partial I_C/\partial V_{CE})\Big|_{V_{BE}} = I_C(\partial \log_e I_C/\partial V_{CE})\Big|_{V_{BE}} \tag{743.13}$$

$$= I_C\cdot\partial \log_e(\alpha_B j_B)/\partial V_{CE}\Big|_{V_{BE}} \tag{743.14}$$

$$= (g_m - g_\mu)\cdot\{-\partial \log_e(\alpha_B j_B)/\partial(qV_{CB}/kT)\}\Big|_{V_{BE}} \tag{743.15}$$

where the bracketed term, in { }, comes from the Early effect and is equal to the Early factor E_A defined in (741.10E) divided by $q|V_{CB}/kT|$. It was characterized by one parameter, the Early Voltage, V_A, in (741.10C). Thus,

$$g_o + 1/(g_\mu^{-1} + g_\pi^{-1}) = \partial I_C/\partial V_{CE}\Big|_{I_B} = I_C/(V_A+V_{CE}) \text{ (npn BJT)} \tag{743.16}$$

$$= -I_C/(V_A-V_{CE}) \text{ (pnp BJT)} \tag{743.17}$$

where V_A is the magnitude of the d.c. Early Voltage. The above equations will give the output conductance, g_o, accurately since the other two terms, the feedthrough conductance g_μ and the input conductance g_π, are both very small.

The **reverse transfer conductance** gives the last of the three equations needed to evaluate the three conductances, g_o, g_μ, and g_π. This is

$$g_\mu = -(\partial I_B/\partial V_{CE})\Big|_{V_{BE}} \tag{743.18}$$

$$= [\exp(-qV_{BE}/kT)]A_E\{\partial[j_E+j_{EB}+j_B(1-\alpha_B)]/\partial V_{CE}\}\Big|_{V_{BE}} \tag{743.19}$$

$$= [\exp(-qV_{BE}/kT)]A_E \cdot \{\partial[j_B(1-\alpha_B)]/\partial V_{CE}\}\Big|_{V_{BE}} \tag{743.20}$$

where use is made of the fact that j_E and j_{EB} are functions of V_{BE} only and do not depend on V_{CE} or V_{CB} when V_{BE} is given or held constant, while j_B and $(1-\alpha_B)$ are dependent on V_{CB} due to collector-base junction space-charge layer thickening or base thickness modulation from the Early effect.

To obtain the four conductance elements of the hybrid-pi equivalent circuit, we first combine (743.20) with (743.14) and get

$$g_\mu = \exp[(-qV_{BE}/kT)]A_E[\partial j_B/\partial V_{CE}] + g_o + g_\mu \tag{743.21}$$

so that

$$g_o = A_E\exp[(-qV_{BE}/kT)][\partial(-j_B)/\partial V_{CE}]\Big|_{V_{BE}} \tag{743.22}$$

$$= -A_E\exp[(-qV_{BE}/kT)](j_B/V_{CE})[\partial\log_e j_B/\partial\log_e V_{CE}]\Big|_{V_{BE}} \tag{743.23}$$

$$= +(I_C/V_{CE})[\partial\log_e j_B/\partial\log_e|V_{CE}|]\Big|_{V_{BE}} \tag{743.24}$$

$$\simeq |I_C/V_{CE}|\cdot E_A/\alpha_B \tag{743.25}$$

$$\simeq |I_C|/[V_A + |V_{CE}|] \tag{743.26}$$

where E_A and V_A are the Early factor and Early voltage defined in (741.10A) and (741.10C) for the CB configuration but are used here for the CE configuration because V_{EB} in the CB and V_{EB} in the CE are related to each other as follow while evaluating these short-circuit output conductances:

$$\partial V_{CE} = \partial V_{CB} + \partial V_{BE} = \partial V_{CB} + 0. \tag{743.27}$$

The result of g_o given by (743.25) is nearly equal to the Early conductance g_a given by (743.10) in the CB configuration. The difference comes from the V_{CB} and V_{CE} voltages in the denominator which is less than 10% if the V_{CB} and V_{CE} are large compared with V_{EB}. The result of $g_o \simeq g_a$ should not be surprising since they are both short-circuit output conductances due to the base-layer thickness modulation or the Early effect.

The feedthrough conductance, g_μ, can be computed from (743.20) if we again make the approximation valid for a high gain and high frequency transistor: $\alpha_B = \text{sech}(X_B/L_B) \approx 1 - X_B^2/(2D_B\tau_B) = 1 - t_B/\tau_B$. Then,

$$\begin{aligned} J_B(1-\alpha_B) &= (qD_BP_B/L_B)\text{ctnh}(X_B/L_B)[1 - \text{sech}(X_B/L_B)] \\ &\approx (qD_BP_B/X_B)(X_B^2/2D_B\tau_B) \\ &= qP_BX_B/2\tau_B. \end{aligned} \qquad (743.28)$$

Using this and (743.24) in (743.20), then,

$$g_\mu = [q(-I_C)/kT][(1-\alpha_B)/\alpha_B][(kT/q)(\partial\log_e X_B/\partial V_{CE})]\Big|_{V_{BE}} \qquad (743.29)$$

$$= (1-\alpha_B)g_o \qquad (743.30)$$

$$= (X_B^2/2D_B\tau_B)g_o = (t_B/\tau_B)g_o. \qquad (743.31)$$

Summary of BJT CEss-Hπ Conductance Elements and Numerical Example

The results of the preceding long algebra are collected below. The derived formulae without approximations are:

$$g_m - g_\mu = (-qI_C/kT) \qquad (743.32)$$

$$g_\pi + g_\mu = (-qI_C/kT)\beta_F^{-1}[1 + E_e] \qquad (743.33)$$

$$g_o = I_C/(V_A+V_{CE}) \qquad (743.34)$$

and

$$g_\mu = (t_B/\tau_B)g_o. \qquad (743.35)$$

The empirical-experimental relationship for the output conductance is

$$(g_\mu^{-1} + g_\pi^{-1})^{-1} + g_o = - I_C/(V_A-V_{CE}). \qquad (743.36)$$

A numerical example is now given. Consider the production Si n/p/n transistor data at 25°C shown in Fig.738.1 at $V_{CE}=+5\text{V}$, $I_C=360\text{mA}$, $I_B=5\text{mA}$, and with Early voltage $V_A=7.4\text{V}$. Assume $\tau_B=1\mu\text{s}$ and $t_B=1\text{ns}$. Then,

$$\beta_F = I_C/I_B = 360/5 = 72 \qquad (743.37)$$

$$g_m = qI_C/kT = 360\text{mA}/25\text{mV} = 14.4\ \text{mho} = 1.4\text{x}10^1\ \text{mho} \qquad (743.38)$$

$$g_o = I_C/(V_A+V_{CE}) = 900/(5+7.4) = 29\ \text{mmho} = 2.9\text{x}10^{-2}\ \text{mho} \qquad (743.39)$$

$$g_\mu = (t_B/\tau_B)g_o = (10^{-9}/10^{-6})\text{x}2.9\text{x}10^{-2} = 2.9\text{x}10^{-5}\ \text{mho} \qquad (743.40)$$

$$g_\pi = (qI_C/kT)\beta_F^{-1}\cdot[1+E_e] - g_\mu = 2.0\text{x}10^{-1}\ \text{mho} \qquad (743.41)$$

$$E_A = V_{CE}/(V_A+V_{CE}) = 5.0/(5.0+7.4) = 0.40 \qquad (743.42)$$

These numerical results show that g_μ can be dropped compared with g_m and g_π in (743.32) and (743.33).

Intrinsic Charge Control Capacitances of CEss-Hπ BJT Model

The charge control capacitances are exactly the same as those derived for the CBss-HTee shown in Fig.741.1(d), C_{eb} and C_{cb}. These can be attached to the eb and cb nodes shown in Fig.743.1(c). Since the collector-base junction is reverse biased, $C_{cb} = C_c + C_{cbt} + C_b$, $\approx C_{cbt} \equiv C_c$ abbreviated with the classical one-subscript, C_c. A small contribution from the Early effect was neglected in the CBss-Tee and CBss-HTee circuits. This is added to C_{cb}, increasing C_c to $C_c + g_{mo} t_B = C_c + g_o t_B$. Generally, $g_o t_B << C_c$ and can be neglected.

Parasitics of the CEss-Hπ BJT Model

The series resistances and the collector-base underlap diode capacitance of the CBss-HTee model of Fig.741.1(e) can also be connected to the internal nodes and terminals of the CEss-Hπ model. The resultant complete CEss-Hπ equivalent circuit model is shown in Fig.743.1(e).

744 Common-Emitter Current Gain, Cutoff Frequency and Bandwidth

One of the most frequently used figure-of-merit of a BJT transistor is the common-emitter short-circuit current gain, $\beta_f(\omega) = -I_c/I_b = -h_{21}$. The frequency at which $|\beta(\omega=\omega_\beta)| = |\beta(\omega=0)|/\sqrt{2}$ or drops by 3db from its low frequency value is known as the **beta cut-off frequency**. The frequency at which beta is unity, $|\beta_f(\omega=\omega_{bw}\equiv\omega_t)| = 1$, is known as the **bandwidth** or **CE unity beta frequency**. These are illustrated by a plot of $\beta_f(\omega)$ versus ω in Fig.744.1.

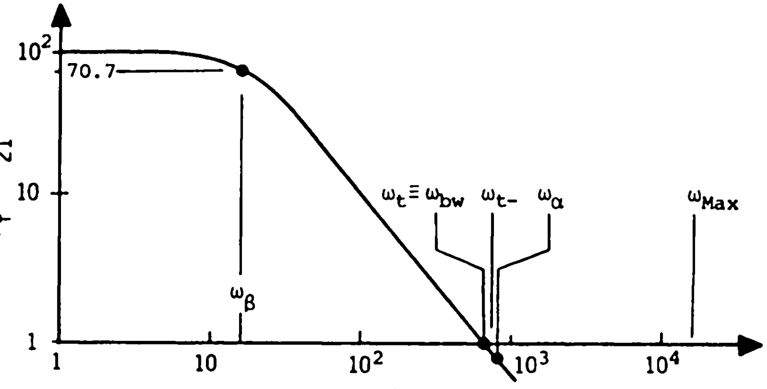


Fig.744.1 Frequency dependence of the CEss short-circuit current gain, $\beta_f(\omega)$.

Derivation of these expressions by straightforward circuit analysis is now given. Instead of the CEss-Hπ model of Fig.743.1(d), the CBss-HTee model shown in Fig.741.1(e) is used because its circuit elements were rigorously derived from the exact 1-d small-signal solution while the elements of the CEss-Hπ model are transformed or indirectly derived involving some approximations. In principle both will give the same results. The simpler algebra using the CEss-Hπ model are given as problems to delineate the approximations.

We shall omit $r_{e'}$, $r_{c'}$ and C_{ul}. Also, $r_{b'}$ will not enter in the short-circuit current gain analysis since it is in series with the constant input current source I_b. Thus, from Fig.741.1(e)

$$I_b = (j\omega C_{cb}+Y_{eb})V_{b'e} + \alpha_f I_e + j\omega C_c V_{b'e} \tag{744.1}$$

$$I_e = - Y_{eb}V_{b'e} \tag{744.2}$$

where

$$Y_{eb} = g_{eb}+j\omega C_{eb} \equiv g_{eb}(1+j\omega/\omega_{eb}) \tag{744.2A}$$

and

$$\alpha_f = \alpha_{fo}/(1+j\omega/\omega_{\alpha'}). \tag{741.32A}$$

Thus, the **CEss forward current transfer ratio** or **CEss forward short-circuit-output current gain** is

$$\beta_f = -I_c/I_b = -(I_b+I_e)/I_b \equiv - h_{21} \tag{744.3}$$

$$= \frac{\alpha_{fo} - (j\omega C_c/g_e)(1+j\omega/\omega_{\alpha'})/(1+j\omega/\omega_{eb})}{1-\alpha_{fo} + j\omega/\omega_{\alpha'} + (j\omega C_c/g_e)(1+j\omega/\omega_{\alpha'})/(1+j\omega/\omega_{eb})} \tag{744.4}$$

$$\propto \frac{\alpha_{fo}}{1-\alpha_{fo}}\;\frac{1 - (j\omega C_c/g_e\alpha_{fo})}{1 + j[\omega/(1-\alpha_{fo})][\omega_{\alpha'}{}^{-1}+(C_c/g_{eb})]} \tag{744.4A}$$

$$\propto \frac{\alpha_{fo}}{1-\alpha_{fo}}\;\frac{1}{1 + j[\omega/(1-\alpha_{fo})][\omega_{\alpha'}{}^{-1}+(C_c/g_{eb})]} \tag{744.4B}$$

$$\equiv \frac{\beta_{fo}}{1 + j\omega/\omega_\beta}\,. \tag{744.5}$$

The reciprocal **beta -3db cutoff frequency** is given by

$$\omega_\beta{}^{-1} = [\omega_{\alpha'}{}^{-1} + (C_c/g_{eb})]/(1-\alpha_{fo}) \tag{744.6}$$

$$= \{(\gamma_{eo}t_B/1.2162) + [(C_e+C_{ebt}+C_c)/g_{eb}] + (\gamma_{eo}X_{CB}/2\theta_{sat})\}\beta_{fo}/\alpha_{fo}$$

$$= \{(t_B/1.2162) + [(C_e+C_{ebt}+C_c)/g_b] + (X_{CB}/2\theta_{sat})\}(\beta_{fo}/\alpha_{bo})$$

$$\propto \{(t_B/1.2162) + [(C_e+C_{ebt}+C_c)/g_b] + (X_{CB}/2\theta_{sat})\}\beta_{fo}. \tag{744.6A}$$

The approximations are: (i) $1+j\omega/\omega_{\alpha'} \simeq 1+j\omega/\omega_{eb}$ is used in (744.4) to give (744.4A) since $\omega_{\alpha'} \simeq \omega_{eb}$; and (ii) $1-\alpha_{fo} << 1$ is used in (744.4A) to give (774.4B), since $\beta_{fo}=\alpha_{fo}/(1-\alpha_{fo}) >> 1$.

The **unity beta frequency** or **bandwidth**, $\omega_{bw}(\equiv\omega_t=2\pi f_t\equiv 2\pi f_{bw})$ is derived by setting $|\beta_f(\omega_{bw})|=1$ in (744.5) which gives

$$\omega_{bw} \equiv \omega_t = 2\pi f_t = \omega_\beta\sqrt{\beta_{fo}{}^2 - 1} \simeq \omega_\beta\beta_{fo} = 1/[\omega_{\alpha'}{}^{-1} + (C_c/g_{eb})] \quad (744.7)$$

$$\neq \omega_{t-} = 2\pi f_{t-} \qquad = 1/[\omega_{\alpha'}{}^{-1} + (C_c r_{b'})] \quad (742.8C)$$

Two key results should be noticed. (i) (744.6A) shows that the beta -3db cutoff frequency is β_{fo} times smaller than ω_α and hence is much easier to obtain experimentally by measuring the CE short-circuit current gain, I_c/I_b, as a function of signal frequency. (ii) The bandwidth, $f_{bw}(\equiv f_t)$, given by (744.7) differs from f_{t-} that appears in the maximum frequency of oscillation formula given by (742.8C) due to C_c/g_{eb} in $f_{bw}(\equiv f_t)$ and $C_c r_{b'}$ in f_{t-}.

A numerical example of a medium speed and medium frequency Si n/p/n transistor is now given to illustrate the magnitude of the various time constants and characteristic frequencies. A discussion of the numerical results is given after the calculations. The following geometry and material parameters will be used.

Emitter:

$A_E=2\mu m\text{x}5\mu m=10^{-7}cm^2$, $N_{EE}=N_{DD+}=10^{19}cm^{-3}$, $\mu_{pe}=66cm^2/V\text{-}s$, $D_E=D_{pe}=1.65cm^2/s$, $X_E=1.0\mu m$, $\tau_E=\tau_{pe}=0.303\mu s$. J_E is large so that $g_{ebt} << g_b$ and emitter-base space-charge recombination can be neglected. But J_E is still in the low injection level range.

Base:

$N_{BB} = N_{AA} = 10^{17}cm^{-3}$, $X_B = 1\mu m$, $\mu_{nb} = 660cm^2/V\text{-}s$, $D_B = D_{nb} = 16.5cm^2/s$, $\mu_{pb} = 200cm^2/V\text{-}s$, $\rho_B = 1/q\mu_{pb}N_{BB} = 0.3125\Omega\text{-}cm$, $R_{B/sq} = \rho_B/X_B = 3125\Omega/sq$, $\tau_B = \tau_{nb} = 0.303\mu s$, $V_A = 75V$.

Collector:

$A_C=2A_E=4\mu m\text{x}5\mu m=2\text{x}10^{-7}cm^2$, $N_{CC} = N_{DD}= 10^{16}cm^{-3}$, $\mu_{nc} = 1000cm^2/V\text{-}s$, $X_C=5\mu m$ to N_{DD++} buried collector, $\theta_{sat}=10^7cm/s$.

Bias and Temperature:

$V_{CB}=-5V$, $J_C \simeq J_E=250A/cm^2$, $I_C=25\mu A$, $kT/q=25mV$, $T\simeq 290K$, $n_i \simeq 10^{10}cm^{-3}$.

The following circuit and device parameters are calculated.

$$t_E = X_E{}^2/2D_E = (1.0\times10^{-4})^2/(2\times1.65)\mathrm{s} \qquad = 3.03\mathrm{ns} \tag{744.8A}$$

$$t_E/\tau_E \qquad = 3.03\mathrm{ns}/303\mathrm{ns} \qquad = 10^{-2} \tag{744.8B}$$

$$g_e/g_b = D_E X_E N_{BB}/D_B X_B N_{EE} = \gamma_{eo}{}^{-1} - 1 = \tag{744.9A}$$

$$= 1.65\times10^{-4}\times10^{17}/16.5\times10^{-4}\times10^{19} \qquad = 10^{-3} \tag{744.9B}$$

$$t_B = X_B{}^2/2D_B = (1.0\times10^{-4})^2/(2\times16.5)\mathrm{s} \qquad = 303\mathrm{ps} \tag{744.10A}$$

$$t_B/\tau_B \qquad = 303\mathrm{ps}/303\mathrm{ns} \qquad = 10^{-3} \tag{744.10B}$$

$$\alpha_{fo} = \alpha_B\gamma_{eo} = (1-t_B/\tau_B)[g_b/(g_b+g_e) \tag{744.11A}$$

$$= (1-10^{-3})/(1+10^{-3}) \qquad = 1 - 2\times10^{-3} \tag{744.11B}$$

$$\beta_{fo} = \alpha_{fo}/(1-\alpha_{fo}) = (1 - 2\times10^{-3})/2\times10^{-3} \qquad = 499 \tag{744.11C}$$

$$g_{mfo} \simeq g_b = qI_C/kT = (25\times10^{-6}/25\times10^{-3})\mathrm{S} \qquad = 1.00\mathrm{mS} \tag{744.12A}$$

$$g_e = (\gamma_{eo}{}^{-1}-1)g_b = 1.0\times10^{-3}\times1.00\mathrm{mS} \qquad = 1.00\mu\mathrm{S} \tag{744.12B}$$

$$C_b = g_b t_B = 1.00\mathrm{mS}\times303\mathrm{ps} \qquad = 303\mathrm{fF} \tag{744.13A}$$

$$C_e = g_e t_E = 1.00\mu\mathrm{S}\times3.03\mathrm{ns} \qquad = 3.03\mathrm{fF} \tag{744.13B}$$

$$X_{CB} = \sqrt{2\varepsilon(V_{CB-bi}-V_{CB})/qN_{MC}} \qquad = 0.905\mu\mathrm{m} \tag{744.14A}$$

$$= \sqrt{2\times11.7\times8.85\times10^{-14}(0.748+5.0)/1.6\times10^{-19}\times9.09\times10^{15}}\mathrm{cm}$$

$$X_{EB} = \sqrt{2\varepsilon(V_{EB-bi}-V_{EB})/qN_{ME}} \qquad = 0.119\mu\mathrm{m} \tag{744.14B}$$

$$= \sqrt{2\times11.7\times8.85\times10^{-14}(1.0-0.0)/1.6\times10^{-19}\times9.09\times10^{16}}\mathrm{cm}$$

$$C_{ebt} = \varepsilon A_E/X_{EB}$$

$$= (11.7\times8.854\times10^{-14}\times10^{-7}/0.119\times10^{-4})\mathrm{F} \qquad = 8.70\mathrm{fF} \tag{744.15A}$$

$$C_{cbt} = (11.7\times8.854\times10^{-14}\times10^{-7}/0.885\times10^{-4})\mathrm{F} \qquad = 1.20\mathrm{fF} \tag{744.15B}$$

$$C_{ui} = C_{cbt}(A_C-A_E)/A_E = C_{cbt}(A_{ui}/A_E) = C_{cbt} \qquad = 1.20\mathrm{fF} \tag{744.15C}$$

$$R_{Bsq} = (q\mu_{pb}N_{BB})^{-1}/X_B$$

$$= (1.6\times10^{-19}\times200\times10^{17}\times10^{-4})^{-1} \qquad = 3125\Omega/\mathrm{sq} \tag{744.16A}$$

$$r_{b'} = (q\mu_{pb}N_{BB})^{-1}(\tfrac{1}{2}W_E/2)/(X_B L_E)$$

$$= R_{Bsq}(W_E/4L_E) = 3125(10^{-4}/4\times5\times10^{-4}) \qquad = 156\Omega \tag{744.16B}$$

$$r_{c'} = A_C(\text{area of } J_C)/[q\mu_{nc}N_{CC}(X_C-X_{CB})] \qquad = 152\Omega \tag{744.16C}$$

$$= A_E/[q\mu_{nc}N_{CC}(X_C-X_{CB})]$$

$$= 10^{-7}/[1.6\times10^{-19}\times1000\times10^{16}(5-0.905)\times10^{-4}]$$

The calculated characteristic frequencies are listed below.

$$\begin{aligned} f_{\alpha b} &= \omega_{\alpha b}/2\pi = 1.2162/(2\pi t_B) \\ &= [2\pi(303\text{ps}/1.2162)]^{-1} = (2\pi \times 249\text{ps})^{-1} &&= 639\text{MHz} \quad (744.17\text{A}) \end{aligned}$$

$$\begin{aligned} f_{\alpha} &= \omega_{\alpha}/2\pi = \omega_{\alpha'}/2\pi\gamma_{eo} \approx \omega_{\alpha}/2\pi \\ &= \{2\pi[\omega_{\alpha b}{}^{-1} + (C_e+C_{ebt})/g_b + (X_{CB}/2\Theta_{sat})]\}^{-1} \\ &= \{2\pi[249 + 3.03 + 8.70 + 4.52]\text{ps}\}^{-1} &&= 600\text{MHz} \quad (744.17\text{B}) \end{aligned}$$

$$\begin{aligned} f_{t-} &= \{2\pi(\omega_{\alpha}{}^{-1} + r_{b'}C_c)\}^{-1} \\ &= \{2\pi(249+3.03+8.70+4.52 + 0.188)\text{ps}\}^{-1} &&= 600\text{MHz} \quad (744.17\text{C}) \\ &\lesssim f_{\alpha} \end{aligned}$$

$$\begin{aligned} f_{Max} &= \sqrt{f_{t-}/8\pi r_{b'}C_c} = \{4\pi\sqrt{t_{t-}r_{b'}C_c}\}^{-1} \\ &\qquad\qquad = \{4\pi\sqrt{267\times 0.188}\text{ps}\}^{-1} &&= 11.2\text{GHz} \quad (744.17\text{D}) \end{aligned}$$

$$\begin{aligned} f_{\beta} &= \omega_{\beta}/2\pi = (1-\gamma_{eo})\{2\pi[\omega_{\alpha'}{}^{-1} + C_c/g_{eb}]\}^{-1} \\ &= 2\times 10^{-3}\{2\pi[249+3.03+8.70+4.52+1.20]\text{ps}\}^{-1} &&= 1.20\text{MHz} \quad (744.17\text{E}) \end{aligned}$$

$$\begin{aligned} f_{bw} &= \omega_{bw}/2\pi = f_t = \omega_t/2\pi = \{2\pi[\omega_{\alpha'}{}^{-1} + (C_c/g_{eb})]\}^{-1} \\ &= \{2\pi[249+3.03+8.70+4.52+1.20]\text{ps}\}^{-1} &&= 597\text{MHz} \quad (744.17\text{F}) \\ = f_t &\neq f_{t-} = 600\text{MHz} && \quad (744.17\text{G}) \end{aligned}$$

Several important conclusions can be drawn from the preceding numerical results which are listed in the following paragraphs. They are useful for designing higher frequency and faster BJTs.

(1) The base diffusion conductance, g_b, dominates over (10^3 larger than) g_e and g_{ebt}. This is a consequence of large d.c. and low-frequency beta.

(2) For the same reason, C_b dominates over C_e and C_{ebt}: $C_b/C_e=100$; $C_b/C_{ebt}=303/8.70=35$ at $V_{EB}=0\text{V}$ and $C_b/C_{ebt}=35/2$ at $V_{EB}=0.75\text{V}$.

(3) $f_{\alpha b}$ or the base diffusion delay, $t_B/1.2162=249\text{ps}$, in this 1-μm base thickness transistor is the principal limitation, giving 249ps delay. Emitter quasi-neutral layer and emitter-base space-charge layer capacitances contribute only $C_e/g_b=3.03\text{ps}$ and $C_{ebt}/g_b = 8.70\text{ps}$ signal delays. Drift delay through the collector-base space-charge layer contributes only 4.52ps. Thus, the first design change to give higher frequency is to thin down the base layer. Thinning the base down 10-times to 0.1μm would reduce the base diffusion delay by 100x to 2.49ps. Then, all the other contributions become important and all the material and geometry parameters must be varied simultaneously to give the highest f_{α} and f_{β}.

An analysis is given in section 755 on the ultrafast and ultrahigh frequency BJTs made by the latest (1990) self-aligned emitter technology.

(4) f_{t-} is not equal to $f_{bw}(\equiv f_t)$ as frequently assumed by BJT authors and development engineers. The difference can be very large if g_{eb} and $1/r_{b'}$ are very different and if t_B is comparable to $r_{b'}C_c$ and C_c/g_{eb}. In this numerical example $g_{eb}^{-1} = 1000\Omega$ and $r_{b'} = 156\Omega$, but t_B is so large (100X larger) that it swamps out the difference. Thinning the base down to $0.1\mu m$ would make f_{t-} very different from $f_t(\equiv f_{bw})$ and using f_t or f_{bw} to compute the f_{Max} would give serious error as indicated in Table 755.1.

(5) f_{Max} (11.2GHz) of this example is nearly 20 times higher than f_α, f_{t-}, and $f_{bw} \equiv f_t$ (all three are about 600MHz) because the $r_{b'}C_c$ is so small (0.187ps) due to the very small collector capacitance C_c (1.2fF). In a medium frequency and medium power BJT, the emitter-base and collector-base junction areas are much larger than assumed ($2\text{x}5\mu m^2$). The f_{Max} would decrease by 10 times to 1GHz if the dimension is increased by 10 times or area by 100 times if the current density is kept constant, or if the current is increased by 100 times to 2.5mA.

(6) If the base layer thickness is reduced 10 times to $0.1\mu m$ to give smaller t_B, larger bandwidth, and higher cutoff frequencies, then $r_{b'}$ would increase which reduces f_{Max}. Thus, to limit the increase of $r_{b'}$, base impurity doping must be increased, base width must be decreased, and better base contact geometry must be used. The self-aligned emitter technology has been successful to reduce $r_{b'}$.

(7) The emitter-base junction space-charge layer capacitance, C_{ebt}, can be reduced to increase the cutoff frequencies but it cannot be reduced indefinitely because the emitter layer must be highly doped and the base layer lowly doped to give a high emitter injection efficiency, γ_{eo}, and high beta, β_{fo}. Heterojunction emitter offers a solution to overcome the conflicting requirements. However, $C_{cbt} = \varepsilon A_E/X_{CB}$ of the collector-base junction space-charge layer can be reduced by decreasing the n-collector layer doping. But, this is also limited since the transit time $X_{CB}/2\theta_{sat}$ would increase. The sum $r_{b'}C_{cbt} + (X_{CB}/2\theta_{sat}) = (r_{b'}\varepsilon A_E/X_{CB}) + (X_{CB}/2\theta_{sat})$ reaches a minimum when the two contributions are equal, i.e., $X_{CB}^2 = 2\theta_{sat}r_{b'}\varepsilon A_E = 2qN_{CC}/[\varepsilon(V_{CB\text{-}bi} - V_{CB})]$ which gives the optimum N_{CC}.

(8) The most easily measured frequency is the -3db beta frequency, $f_\beta = 1.2$MHz due to the very high d.c. beta. Even if beta is dropped from 499 to 100, this cutoff frequency is still only 6.0MHz and easily measurable. To determine $f_{bw}(\equiv f_t)$, the output short-circuit CE current gain, I_c/I_b, can be measured as a function of signal frequency to a frequency several times f_β. Then, f_{bw} can be computed from $\beta_{fo}f_\beta$. The bandwidth, f_{bw}, is one of the performance parameters almost always given in reports on the latest developments of ultrahigh frequency (100GHz) BJTs and ultrafast (< 10ps) BJTs. It is customarily labeled f_t or f_T and sometimes mistakenly used as f_{t-} to compute f_{Max}.

750 LARGE-SIGNAL SWITCHING CHARACTERISTICS OF BJT

The BJT transistor is highly nonlinear. A linearized circuit model from charge-control analysis would not give a very accurate prediction of the large signal transient response of the BJT when it is switched by even a simple change of input voltage or current. However, charge-control analysis is the only method that can give an analytical solution which covers the entire switching transient by piecewise linear approximation of each phase of the transient. Exact analytical solutions can be obtained only for some parts of the switching transient and will be used to illustrate the accuracy of the charge-control analysis.

The charge-control method was first introduced into college classroom by the SEEC (Semiconductor Electronics Education Committee) which held its summer workshop at MIT in 1960 and published a 7-volume paperback/softcover textbook series on the material physics, device theory, and circuit applications of bipolar transistors [750.1]-[750.7]. This series is still the most complete and accurate on device and circuit physics among current textbooks on the bipolar transistor. Nevertheless, some updating, extension, revision, and minor correction on interpretations are necessary on BJT switching transient analyses which will be given in the following sections.

The 1960-1966 and current affiliations of the authors listed in the 1983 IEEE Membership Directory are shown in parenthesis.

[750.1] R.B.Adler (MIT), A.C.Smith (MIT) and R.L.Longini (CIT → CMU), **Introduction to Semiconductor Physics**, Semiconductor Electronics Education Committee, Volume 1, John Wiley & Sons, New York (1964).
[750.2] P.E.Gray (MIT), D.DeWitt (IBM), A.R.Boothroyd (Queens → Carleton U.), and J.F.Gibbons (Stanford), **Physical Electronics and Circuit Models of Transistors**, SEEC, Volume 2, Wiley (1964).
[750.3] C.L.Searle (MIT), A.R.Boothroyd (Queens → Carleton U.), E.J.Angelo,Jr. (Brooklyn → Bell-Labs.), P.E.Gray (MIT), and D.O.Pederson (Berkeley), **Elementary Circuit Properties of Transistors**, SEEC, Volume 3, Wiley (1964).
[750.4] R.D.Thornton (MIT), D.DeWitt (IBM), E.R.Chenette (Minnesota→Florida), P.E.Gray (MIT), **Characteristics and Limitations of Transistors**, SEEC, Volume 4, Wiley (1966).
[750.5] R.D.Thornton (MIT), C.L.Searle (MIT), D.O.Pederson (Berkeley), R.B.Adler (MIT), and E.J.Angelo,Jr. (Brooklyn→Bell-Labs.), **Multistage Transistor Circuits**, SEEC, Volume 5, Wiley (1965).
[750.6] J.N.Harris (MIT-Lincoln Lab.), P.E.Gray (MIT), and C.L.Searle (MIT), **Digital Transistor Circuits**, SEEC, Volume 6, Wiley (1966).
[750.7] R.D.Thornton (MIT), J.G.Linvill (Stanford), E.R.Chenette (Minnesota→Florida), H.L.Ablin (Nebraska→N.Arizona-St.), J.N.Harris (MIT-Lincoln Lab.), A.R.Boothroyd (Queens→Carleton U.), J.Willis (UCLA), and C.L.Searle (MIT), **Handbook of Basic Transistor Circuits and Measurements**, SEEC, Volume 7, Wiley (1966).

The description and the mathematics of the charge-control method approach employed here emphasizes the physical origin, device physics, and terminal waveforms of the space-time dependences of the transport (conductance) and storage (capacitance) of electrons and holes in the various regions of the transistor. A consistent set of symbols is used which conforms with the IEEE Standard: Greek characters, τ and μ, are used for fundamental material constants, such as τ_B, τ_{EB}, τ_E, τ_{CB}, τ_C for the recombination-generation lifetimes; English characters are used for derived transport time constants, such as $t_B = X_B^2/2D_B$ for the minority carrier transit time through the base layer of thickness X_B due to diffusion delay, $t_{CB\text{-}sat} = X_{CB}/\theta_{sat}$ for the carrier transit time through the collector-base junction space-charge layer due to drift in an electric field, and t_I and t_{II} respectively for phases I and II in a two-phase analysis of a transient waveform.

In the next section, 751, the two governing differential equations are systematically derived and illustrated by examples. The first equation is the exact partial differential equation of the space-time dependent variable, **volume concentration of carrier**, p(x,y,z,t). The second equation is the exact charge-control ordinary differential equation of the time dependent variable, **areal concentration of carrier or charge**, $q_P(t)$, defined by $q\int p(x,y,z,t)dx$ in one-dimensional problem. The physical descriptions and mathematical analyses of switching transient waveforms are given in the next two sections, 752 for the common-base configuration, and 753 for the common-emitter configuration. A summary of the CB and CE transients is given in section 754. Device structures and material parameters to speed up the BJT switching via novel and new technology are described in section 755 which contains a table (Table 755.1) that compares the performance of four BJTs with 1-micron and 0.1-micron linewidth. Digital circuit applications of the BJT are described in the following sections, 76n.

751 The Diffusion and Charge-Control Equations

The two one-dimensional exact equations are derived in this section. They are the partial differential equation of the space-time dependent carrier volume concentration and the charge-control equation of the time-dependent carrier areal concentration. They were previously derived for switching transient analyses of the two-terminal p/n junction diodes in (551.1)-(551.7A). They are extended here to three-terminal devices such as the BJT.

From the Shockley equations (350.2) and (350.4), the one-dimensional continuity and current equations are

$$q\partial p(x,t)/\partial t = -\partial j_P(x,t)/\partial x + q[g_P(x,t) - r_P(x,t)] \qquad (751.1)$$

and

$$j_P(x,t) = -qD_p\partial p(x,t)/\partial x + q\mu_p E_x(x,t)p(x,t) \qquad (751.2)$$

$$j_P(x,t) \approx -qD_p\partial p(x,t)/\partial x + q\mu_p E_x p(x,t) \qquad \text{Constant } E_x \quad (751.2A)$$

$$j_P(x,t) \approx -qD_p\partial p(x,t)/\partial x \qquad \text{Zero } E_x. \quad (751.2B)$$

Using (751.2), the hole current density, $j_P(x,t)$, can be eliminated in (751.1) to give the **1-d space-time partial differential equations** for the hole concentration, p(x,t),

$$\partial p/\partial t = D_p\partial^2 p/\partial x^2 + (g_P - r_P) + \mu_p[\partial(pE_x)/\partial x] \qquad \textbf{Exact } \mathbf{E_x} \quad (751.3)$$
$$\partial p/\partial t \approx D_p\partial^2 p/\partial x^2 + (g_P - r_P) + \mu_p E_x[\partial p/\partial x] \qquad \textbf{Constant } \mathbf{E_x} \quad (751.3A)$$
$$\partial p/\partial t \approx D_p\partial^2 p/\partial x^2 + (g_P - r_P) \qquad \textbf{Zero } \mathbf{E_x} \quad (751.3B)$$

They can be reduced to the following equations if the recombination-generation processes can be represented by a **recombination lifetime**, τ_p, (see sections 37n)

$$\partial p/\partial t = D_p\partial^2 p/\partial x^2 - (p-P_E)/\tau_p + \mu_p[\partial(pE_x)/\partial x] \qquad \textbf{Exact } \mathbf{E_x} \quad (751.4)$$
$$\partial p/\partial t \approx D_p\partial^2 p/\partial x^2 - (p-P_E)/\tau_p + \mu_p[E_x\partial p/\partial x] \qquad \textbf{Constant } \mathbf{E_x} \quad (751.4A)$$
$$\partial p/\partial t = D_p\partial^2 p/\partial x^2 - (p-P_E)/\tau_p \qquad \textbf{Zero } \mathbf{E_x} \quad (751.4B)$$

The second differential equation is the **charge-control analysis equation.** It is obtained by a simple integration of the exact 1-d space-time partial differential equations (751.3)-(751.3B) or (751.4)-(751.4B) over the thickness of the layer. The dependent variable in the 1-d charge-control analysis is the **excess carrier charge** over the equilibrium carrier charge enclosed in the layer bounded by the planes $x=x_1$ and $x=x_2$ and defined by

$$q_P(x_1,x_2,t) = \int_{x_1}^{x_2} q[p(x,t) - P_E]dx. \qquad (751.5)$$

In analyses of minority carrier transport, the equilibrium charge, P_E, is frequently neglected because it is time-independent and very small. Using this definition, the **charge-control** ordinary **differential equation** can be obtained by integrating the continuity equation (751.1) over the 1-d space of interest. It is given by

$$\int_{x_1}^{x_2} q\partial p(x,t)/\partial t = \partial q_P(x,x_1,t)/\partial t$$
$$= -\int_{x_1}^{x_2}[\partial j_P(x,t)/\partial x]dx + \int_{x_1}^{x_2} q[g_P(x,t) - r_P(x,t)]dx$$
$$= + j_P(x_1,t) - j_P(x_2,t) + \int_{x_1}^{x_2} q[g_P(x,t) - r_P(x,t)]dx \qquad (751.6)$$
$$= + j_P(x_1,t) - j_P(x_2,t) + \int_{x_1}^{x_2} q\{[P_E - p(x,t)]/\tau_p(x,t)\}dx$$
$$= + j_P(x_1,t) - j_P(x_2,t) - q_P(x_1,x_2,t)/\tau_p. \qquad (751.6A)$$

Equations (751.6) and (751.6A) are the time-dependent **charge-control ordinary differential equations** for the dependent charge variable $q_P(x_1,x_2,t)$ or $q_P(t)$ in a selected volume enclosed by the planes $x=x_1$ and $x=x_2$. Equation

(751.6A) assumes the existence of a constant recombination lifetime. Both equations are a statement of **charge conservation** or **particle conservation**: the rate of increase of the hole concentration in the volume element or slice between the x_1 and x_2 planes is equal to the rate of generation minus the rate of recombination of holes in this volume, plus the holes flowing into this volume from the plane located at x_1 and minus the holes flowing out of this volume from the plane located at x_2. Equation (751.6A) is rearranged as follows

$$\partial q_P(x_1,x_2,t)/\partial t + q_P(x_1,x_2,t)/\tau_p = j_P(x_1,t) - j_P(x_2,t) \qquad (751.6B)$$

or

$$\partial q_P(t)/\partial t + q_P(t)/\tau_p = j_P(x_1,t) - j_P(x_2,t) \qquad (751.6C)$$

which provides new insights on the solutions of several graphically illustrated examples given in the following paragraphs.

Charge-Control Equations of a General Slab

Figure 751.1(a) gives the cross-sectional view of a **three-terminal** semiconductor plate between the planes x_1 and x_2. It shows the currents flowing into and out of the plate based on the 1-d charge-control equation (751.6B) or (751.6C). The positive reference directions of the terminal and internal currents are shown by the arrows. Applying the Kirchoff current law, $j_P(x_1,t)-j_P(x_2,t)=j_P(x,t)$ to (751.6B), the charge-control equation of this 3-terminal volume element is then

$$\partial q_P(x_1,x_2,t)/\partial t + q_P(x_1,x_2,t)/\tau_p = j_P(x_1,t) - j_P(x_2,t) \qquad (751.6B)$$

$$= j_P(x,t). \qquad (751.6D)$$

Charge-Control Equations of the Entire BJT

Figure 751.1(b) shows the cross-sectional view of a p/n/p BJT whose charge-control equation is given by (751.6B) with $x_1=X_E$, $x_2=X_C$, $j_P(x_1,t)=j_P(X_E,t)=j_E(t)$, and $j_P(X_C,t)=j_P(x_T,t)=j_C(t)$. The Kirchoff current law is $j_P(x,t) = j_E(t) - j_C(t) = j_B(t)$.

$$\partial q_P(x_1,x_2,t)/\partial t + q_P(x_1,x_2,t)/\tau_p = j_P(x_1,t) - j_P(x_2,t) \qquad (751.6B)$$

$$\partial q_P(t)/\partial t + q_P(t)/\tau_p = j_B(t) - j_C(t) \qquad (751.7A)$$

$$= j_B(t). \qquad (751.7B)$$

Charge-Control Equations of the Quasi-Neutral Base Layer

Figure 751.1(c) shows the application of the charge-control equations to the quasi-neutral base layer. Using $x_1=0_B$ and $x_2=X_B$, then

$$\partial q_P(x_1,x_2,t)/\partial t + q_P(x_1,x_2,t)/\tau_p = j_P(x_1,t) - j_P(x_2,t) \qquad (751.6B)$$

$$\partial q_B/\partial t + q_B/\tau_B = j_P(0_B,t) - j_P(X_B,t) \qquad (751.8A)$$

$$= j_B(t). \qquad (751.8B)$$

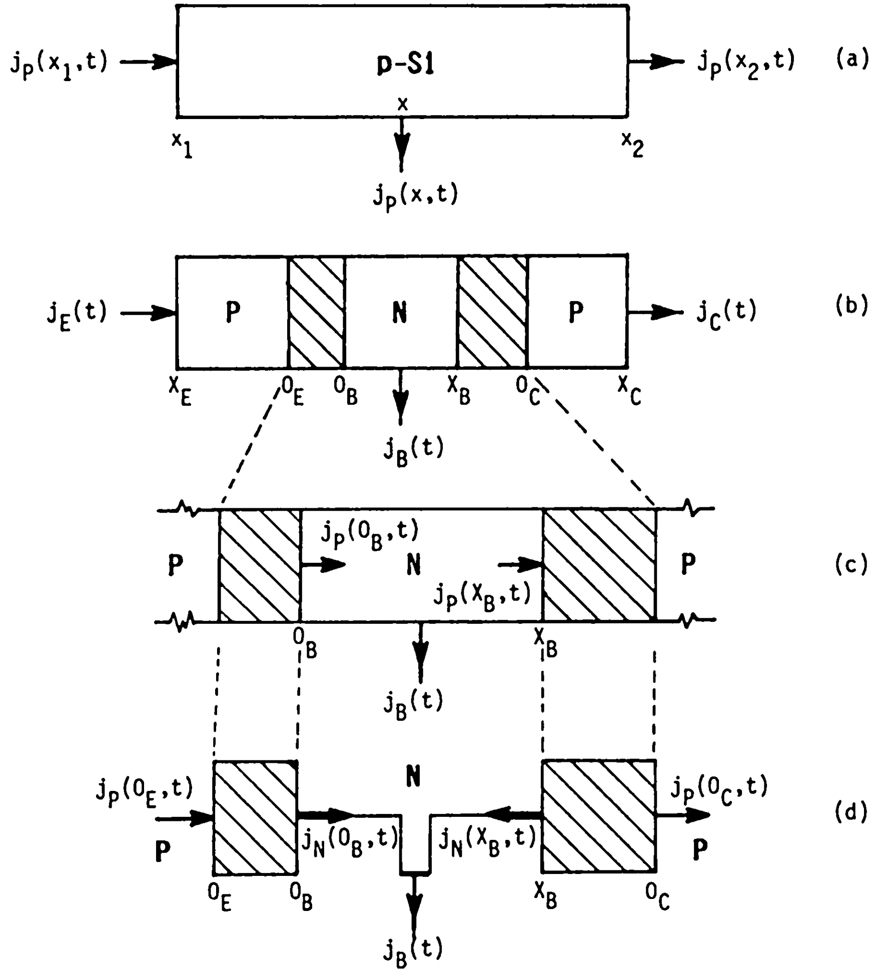

jp(x1,t)
p-Si
x
jp(x2,t)
(a)
x1
x2
jp(x,t)
jE(t)
P
N
P
jC(t)
(b)
XE
0E
0B
XB
0C
XC
jB(t)
P
jp(0B,t)
N
jp(XB,t)
P
(c)
0B
XB
jB(t)
jp(0E,t)
P
N
jN(0B,t)
jN(XB,t)
jp(0C,t)
P
(d)
0E
0B
XB
0C
jB(t)

Fig.751.1 Examples of using and applying the charge-control equation. (a) General slab. (b) Entire BJT. (c) Quasi-neutral base layer. (d) Emitter-base and collector-base space-charge layers.

Charge-Control Equations of the Space-Charge Layers

Figure 751.1(d) shows the application of the charge-control equations to the emitter-base and collector-base space-charge layers. They are the **majority-carrier** charge-control equations in contrast to the base layer where (751.8B) is the **minority-carrier** charge-control or stored-charge equation. Thus, these are multi-phase approximations during which the minority carrier current and recombination-generation in the space-charge layers can be neglected.

Emitter-Base Space-Charge Layer

$$\partial q_P(x_1,x_2,t)/\partial t + q_P(x_1,x_2,t)/\tau_p = + j_P(x_1,t) - j_P(x_2,t) \qquad (751.6B)$$

$$\partial q_P(0_E,0_B,t)/\partial t + 0 = + j_P(0_E,t) - 0$$

$$\partial q_{EB}/\partial t = + j_E(t). \qquad (751.9)$$

Collector-Base Space-Charge Layer

$$\partial q_{CB}/\partial t + 0 = - j_C(t). \qquad (751.10)$$

Complete Base-Charge-Control Equations

If we retain only charge storage and electron-hole recombination in the quasi-neutral base layer, and neglect all the rest (the diffusion-drift-recombination-generation processes in the following layers: emitter, emitter-base space-charge, collector-base space-charge, and collector), then a set of general base-charge-control equations can be derived. This set is the most popular for introducing college sophomores and juniors to BJT switching transient analysis. It relates the terminal currents (I_C, I_B, and I_E) to the stored charge in the quasi-neutral base layer (q_B) in a BJT transistor. The equations can be derived from the foregoing results. In order to delineate and focus on the features of the charge-control method, only the minority carrier diffusion and recombination currents in the base layer are considered. Minority carrier drift current due to an electric field in the base layer is excluded. These excluded currents and stored charges can be added later if the linear approximation is still valid. Neglecting all but the base current, then,

$$j_P(0_B,t) = j_E(t) \qquad (751.11A)$$

and

$$j_P(X_B,t) = j_C(t). \qquad (751.11B)$$

The common-base configuration is used to illustrate this derivation without loss of generality. Since linear charge-current relationship is assumed, the stored base charges injected and extracted by each applied terminal or junction voltage are additive. Thus, the current response to an applied voltage $v_{EB}(t)$ can be added to the current response to an applied voltage $v_{CB}(t)$ because **responses are additive in a linear system**. This linearity between stored charge and current was also the fundamental basis of the original Ebers-Moll equations. But it must be modified in the SNS equations to account for nonlinear recombination-generation of electrons and holes in the space-charge layers. A similar modification must be made to include the time-dependence of the majority carrier charges stored in the space-charge layers. A simple extension can also be made to include the nonlinear high injection level effect (see the d.c. case in section 738C), which is especially simple if recombination in the base is zero as used by Gummel and Poon to develop the widely used Gummel-Poon models.

To focus, let us use the linearity property so that we can study the response of an n/p/n BJT to an emitter-base voltage excitation. Thus, let $v_{CB}(t)=0$ and

$v_{EB}(t)>0$ so that the emitter-base junction is forward biased and injects the minority carriers (holes) into the base. The excess minority carrier (holes) charge in the base layer will be denoted by $q_P(t) = q_{BF}(t)$ where subscript F signifies its origin, i.e. the forward circuit configuration with a forward voltage applied to the emitter-base junction. The double subscript notation is the first of a series of slight modifications and extensions of the original MIT SEEC notations. In particular, the new notation is easier to remember because it is based not only on device physics but also on the IEEE Symbols Standard of using Greek letters for fundamental material constants and parameters, and English letters for derived and circuit parameters. The subscript F allows us to use the subscript R to label the excitation and responses in the reversely connected (output-input interchanged) circuit configuration, i.e., $v_{EB}(t)=0$ and $v_{CB}(t)>0$ or the collector-base junction is forward biased and injecting. This expedites writing down the complete equations using the additive law owing to linearity of the system, without further algebra. We shall still use the areal current density, j_P (A/cm^2). Thus, q_P is the areal charge density (Coulomb/cm^2) instead of the total charge which has been used in the SEEC series and the Berkeley volumes. This choice helps to apply device physics to the geometrical partitioning methodology used in earlier parts of this chapter to treat the three-dimensional effects on the initial phase of the transient waveforms, such as the effects of the underlap collector-base n/p diode, the lateral spreading base resistance, and stray capacitances and inductances. The total current is just the areal current densities multiplied by the corresponding perpendicular cross-sectional areas through which the currents flow.

The first equation is (751.8B) since it is generally valid. It is

$$\partial q_{BF}/\partial t \ + \ q_{BF}/\tau_{BF} \ = + \ j_{BF}(t). \qquad (751.12A)$$

The second equation is derived by **defining a characteristic time constant**, t_{BF}, so that the collector current can be computed from the stored base charge, $q_{BF}(t)$. There are no approximations, no logical reasons, and no physics base, that are required to define t_{BF} as implied by the SEEC authors and some later followers. The only test is whether it predicts experiments well and whether it agrees with the asymptotic form of the exact solution. The defining equation is

$$j_{CP}(t) = q_{BF}(t)/t_{BF} \quad \{\textbf{Definition of } t_{BF}.\} \qquad (751.12B)$$

The explicit formula of t_{BF} can be derived. It will be shown shortly that it is the diffusion delay time through the base layer, $t_{BF}=X_B^2/2D_B$.

Using Kirchoff's current law, the emitter current can be obtained by adding the base and collector currents, (751.8A) and (751.8B). This third equation is

$$\begin{aligned} j_{EP}(t) &= j_{BP}(t) + j_{CP}(t) \\ &= \partial q_{BP}(t)/\partial t + q_{PB}(t)/\tau_{BP} + q_{BP}(t)/t_{BP}. \end{aligned} \qquad (751.12C)$$

A similar set of three equations can be obtained for the reverse configuration, i.e. $v_{EB}(t)=0$ and $v_{CB}(t)>0$ = forward bias. They can be written down without algebra because of the modified new notation of double subscript. Replacing F by R in (751.12A) to (751.12C), the three base-charge/current equations are obtained.

Invoking the linearity or additive property: $j_B(t) = j_{BF}(t)+j_{BR}(t)$, $j_C(t) = j_{CF}(t)+j_{CR}(t)$, and $j_E(t) = j_{ER}(t)+j_{EF}(t)$, then, the complete charge-control equations for the quasi-neutral p-type base of p/n/p BJTs are

$$j_B(t) = \partial q_{BF}/\partial t + q_{BF}/\tau_{BF} \quad + \quad \partial q_{BR}/\partial t + q_{BR}/\tau_{BR} \qquad (751.13A)$$

$$j_C(t) = q_{BF}/t_{BF} \quad - \quad (\partial q_{BR}/\partial t + q_{BR}/\tau_{BR} + q_{BR}/t_{BR}) \qquad (751.13B)$$

$$j_E(t) = \partial q_{BF}/\partial t + q_{BF}/\tau_{BF} + q_{BF}/t_{BF} \quad - \quad q_{BR}/t_{BR}. \qquad (751.13C)$$

These three linear first-order coupled ordinary differential equations will be solved in the following sections to give the transient response of a BJT connected in the CB and CE configurations.

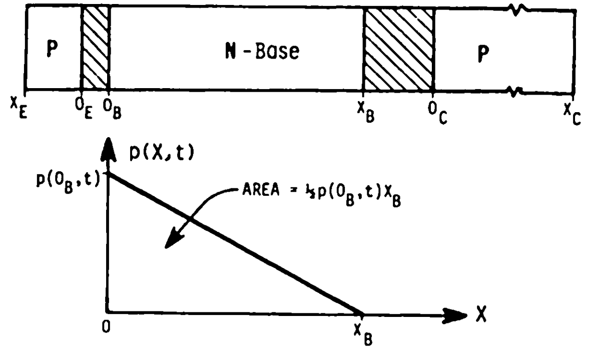


Fig.751.2 The minority carrier (hole) distribution in the quasi-neutral base.

The Base Transport Time Parameters, t_{BF} and t_{BR}

The characteristic time parameter, t_{BF}, is the minority carrier diffusion delay time through the base layer of thickness $X_B=x_2-x_1$. This can be proved using the charge distribution in the base shown in Fig.751.2 for the conditions of $v_{EB}(t)>0$ and $v_{CB}=0$ so $q_{BR}=0$. From (751.12B) and using Fig.751.2, we have

$$j_{CF}(t) = q_{BF}(t)/t_{BF} \quad = qp(0_B,t)\cdot(X_B/2)/t_{BF} \qquad (751.14A)$$

and

$$j_{CF}(t) = -qD_B\partial p(x,t)/\partial x\Big|_{X_B} = qp(0_B,t)\cdot D_B/X_B \qquad (751.14B)$$

which can be combined to eliminate $p(0_B,t)$ to give

$$t_{BF} = X_B^2/2D_B \qquad (751.14C)$$

This is precisely the minority carrier (hole) diffusion transit time through the base layer. A similar expression can be derived for t_{BR} in the reverse direction. t_{BR} differs from t_{BF} only in **graded base** BJT which has a built-in electric field in the base layer from position variation of the base-dopant impurity concentration or the chemical composition (such as the Ge atomic fraction, x, in the smaller energy-gap Ge_xSi_{1-x} base layer of the heterostructure BJT described in sections 77n). If the built-in electric field aids minority carrier diffusion from the emitter to the collector, then it retards in the reverse direction. The aiding drift field in the forward connnection gives a multiplicity factor, $\eta_F > 1$, which modifies (751.14C) to give

$$t_{BF} = X_B^2/2\eta_F D_B. \qquad (751.14D)$$

If the base impurity concentration changes by $10^3 \approx \exp(7)$ from the emitter-base to the base-collector junction, the built-in field gives a potential drop of 7kT/q. This gives an aiding factor of $\eta_F = 7$ or an effective diffusivity $7D_B$. This built-in drift field would oppose minority carrier diffusion in the reverse direction (injected from the collector towards the emitter), giving $\eta_R < 1$ or $\eta_R = 1/7$. Thus,

$$t_{BR} = X_B^2/2\eta_R D_B. \qquad (751.14E)$$

At high forward bias or **high injection level**, the large major carrier density gradient in the base layer will produce an **aiding electric field** to accelerate minority carriers in either the forward or reverse connection. This majority-carrier electric field will wipe out the effect of the built-in electric field. In the high- level limit, $p(x,t)-P_E >> 10N_E$ in n-type quasi-neutral base, and the diffusivity is given by the ambipolar diffusivity defined by $D_H = 2D_nD_p/(D_n+D_p)$. Thus, $\eta_F D_B$ and $\eta_R D_B$ are replaced by D_H. Then, $D_H = D_p$ if $D_n = D_p$, and $D_H = 2D_p$ if $D_n >> D_p$.

Relating the Charge-Control and Ebers-Moll Parameters

The charge-control parameters or characteristic times can be related to the original or extended Ebers-Moll d.c. parameters of the BJT. Set $\partial/\partial t = 0$ for the d.c. steady-state condition, then from (751.13A)–(751.13C), we have

$$\begin{aligned}
J_B &= +(\tau_{BF}^{-1} \qquad\qquad)Q_{BF} + (\tau_{BR}^{-1} + \qquad\quad)Q_{BR} && (751.15A)\\
J_C &= +(\qquad\quad + t_{BF}^{-1})Q_{BF} - (\tau_{BR}^{-1} + t_{BR}^{-1})Q_{BR} && (751.15B)\\
J_E &= +(\tau_{BF}^{-1} + t_{BF}^{-1})Q_{BF} - (\qquad\qquad t_{BR}^{-1})Q_{BR}. && (751.15C)
\end{aligned}$$

Using the present current reference directions for a p/n/p transistor, the EB equations from (734.3) and (734.4) are

$$\begin{aligned}
J_B &= J_E - J_C && (751.16A)\\
-J_C &= -\alpha_F J_{ES}[\exp(qV_{EB}/kT)-1] + J_{CS}[\exp(qV_{CB}/kT)-1] && (751.16B)\\
J_E &= +J_{ES}[\exp(qV_{EB}/kT)-1] - \alpha_R J_{CS}[\exp(qV_{CB}/kT)-1]. && (751.16C)
\end{aligned}$$

There is no need to run through any algebra, as the SEEC series and the recent Berkeley volumes have done. A mere term by term comparison of the charge-control equations (751.15A) to (751.15C) with the Ebers-Moll equations (751.16A) to (751.16C) immediately gives

$$Q_{BF} = (\tau_{BF}^{-1} + t_{BF}^{-1})^{-1} J_{ES}[\exp(qV_{EB}/kT) - 1] \tag{751.17}$$

$$= t_{BF}\alpha_F J_{ES}[\exp(qV_{EB}/kT) - 1] \tag{751.17A}$$

$$= Q_{BFO}[\exp(qV_{EB}/kT) - 1] \tag{751.17B}$$

where

$$Q_{BFO} = (\tau_{BF}^{-1} + t_{BF}^{-1})^{-1} J_{ES} = t_{BF}\alpha_F J_{ES} \tag{751.17C}$$

$$J_{ES} = Q_{BFO}/t_{BF}\alpha_F \tag{751.17D}$$

$$\alpha_F = t_{BF}^{-1}(\tau_{BF}^{-1} + t_{BF}^{-1})^{-1} = (1 + t_{BF}/\tau_{BF})^{-1} \tag{751.18}$$

$$= [1+(X_B^2/2D_B\tau_{BF})]^{-1} \tag{751.18A}$$

$$\beta_F = \alpha_F/(1-\alpha_F) = \tau_{BF}/t_{BF} = 2D_B\tau_{BF}/X_B^2 \tag{751.19}$$

and

$$Q_{BR} = (\tau_{BR}^{-1} + t_{RF}^{-1})^{-1} J_{CS}[\exp(qV_{CB}/kT) - 1] \tag{751.20}$$

$$= t_{BR}\alpha_R J_{CS}[\exp(qV_{CB}/kT) - 1] \tag{751.20A}$$

$$= Q_{BRO}[\exp(qV_{EB}/kT) - 1] \tag{751.20B}$$

where

$$Q_{BRO} = (\tau_{BR}^{-1} + t_{BR}^{-1})^{-1} J_{CS} = t_{BR}\alpha_R J_{CS} \tag{751.20C}$$

$$J_{CS} = Q_{BRO}/t_{BR}\alpha_R \tag{751.20D}$$

$$\alpha_R = t_{RF}^{-1}(\tau_{BR}^{-1} + t_{BR}^{-1})^{-1} = (1 + t_{BR}/\tau_{BR})^{-1} \tag{751.21}$$

$$= [1+(X_B^2/2D_B\tau_{BR})]^{-1} \tag{751.21A}$$

$$\beta_R = \alpha_R/(1-\alpha_R) = \tau_{BR}/t_{BR} = 2D_B\tau_{BR}/X_B^2. \tag{751.22}$$

752 Common-Base Large-Signal BJT Switching Transients

We now describe the physics and give the derivation of the transient equations of the short-circuit Common-Base Large-Signal switching transients (CBLS-st) in a p/n/p BJT after a current step is applied to the emitter lead. The analysis of the transients due to a voltage step applied to the emitter-base terminals is more complicated which tends to obscure the switching physics.

The circuit diagram of the p/n/p BJT is shown in Fig.752.1(a). In order to avoid double negative due to negative terminal currents, the reference directions of the base and collector current are reversed from the convention, i.e., positive charges flowing out of the base and collector terminals. The reference direction of the emitter current follows the convention, i.e., positive charges flowing into the emitter terminal. Thus, $i_E(t)$, $i_B(t)$ and $i_C(t)$ are all positive during the entire turn-on and unforced turn-off transients. Unforced or natural turn-off refers to short-

or open-circuiting the emitter input after the emitter and collector currents have reached steady-state. In contrast, a forced turn-off applies a reverse current or reverse voltage through a resistance to the emitter-base terminals to speed up the turn-off transient. The collector-base junction has a large d.c. reverse bias voltage.

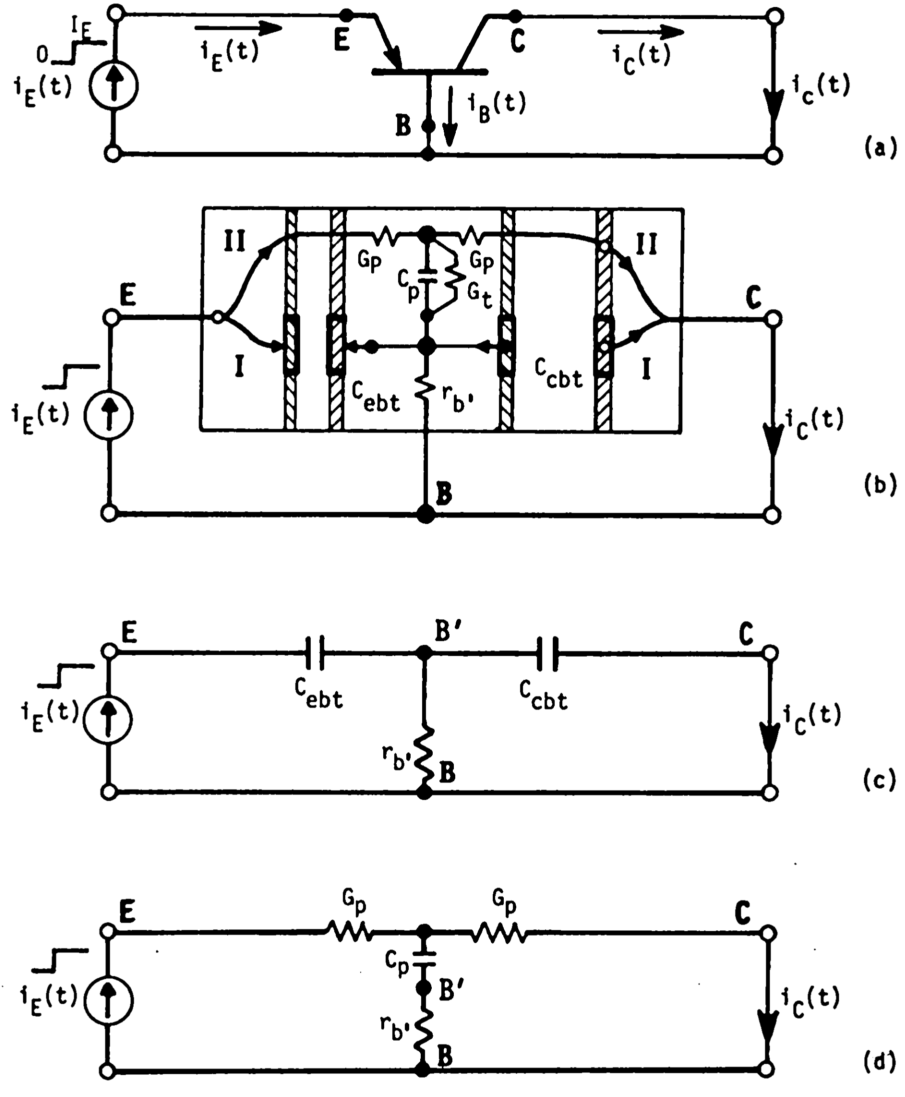


Fig.752.1 Common-base large-signal short-circuit turn-on transient by an emitter current step in a p/n/p BJT. (a) The circuit diagram. (b) The physical circuit with phase I and II current paths. (c) The phase I circuit diagram. (d) The phase II circuit diagram.

The physical circuit diagram is shown in Fig.752.1(b). It is the key to understand the basic physics and to simplify the analysis. In this figure, the base layer is represented by one T-section whose elements are the minority carrier (hole) drift-diffusion conductance G_p and storage capacitance C_p, and the electron-hole recombination conductance G_t at the traps. The T section represents the diffusion-drift delay and the small recombination loss in the base layer. The majority-carrier elements are also shown: the emitter-base (C_{ebt}) and collector-base (C_{cbt}) space-charge layer capacitances and the lateral base resistance ($r_{b'}$). This figure also shows the directions of movement of the electrons (dots) and holes (circles) along the two current paths, I and II. Path I gives the fast initial phase (phase I) due to charging of the space-charge layer capacitances by the majority carriers (electrons) through the base resistance, $r_{b'}$ (and by holes in the quasi-neutral p-type emitter and collector layers to charge up the other plate of the two space-charge layer capacitances). Path II gives the slower second phase (phase II) due to minority carrier (holes) diffusion delay in the base layer. When the delays along the two paths are very different, the two phases can be separated and represented by two equivalent circuits: Fig.752.1(c) for path I of the faster phase I, and Fig.752.1(d) for path II of the slower phase II. However, in state-of-the-art picosecond BJTs, the two time constants may be comparable and the transient waveforms cannot be decomposed into two phases and represented by two equivalent circuits.

Common-Base Turn-On Transients in BJT

The turn-on transient waveforms are shown in Figs.752.2(a) to (d). Figure 752.2(a) is the input emitter current waveform. Figure 752.2(b) is the emitter-base junction voltage waveform, $v_{EB'}(t)$, which differs from the emitter-base terminal voltage waveform, $v_{EB}(t)$, due to the time-varying base current flowing through the base resistance, $r_{b'}$. Figure 752.2(c) is the collector-base junction voltage waveform, $v_{B'B}(t)$, which is also the voltage drop across the base resistance. Figure 752.2(d) gives the collector and base current waveforms which are complements since $i_B(t)+i_C(t)=i_E(t)=I_E=1$ if the emitter is driven by a unit current step or if the solutions are normalized to the magnitude of the current step, I_E. The equations of the waveforms will be derived from the physical and circuit models given respectively in Figs.752.1(b), (c) and (d).

Notice that the turn-on collector current transient shown in Fig.752.2(d) can be visibly divided into two time phases. The sharpness of the time boundary is determined by two of the three time constants of this system. There are three times constants since there are three elements: the stored mobile majority carriers in the quasi-neutral emitter and base layers represented by the space-charge layer capacitances of the emitter-base and collector-base junctions which are charged and discharged through the spreading base resistance (giving 2 time constants), and the minority carrier charge stored in the quasi-neutral base layer (giving 1 time constant). A fourth component and time constant would be present if minority carrier stored in the quasi-neutral emitter layer is not negligible.

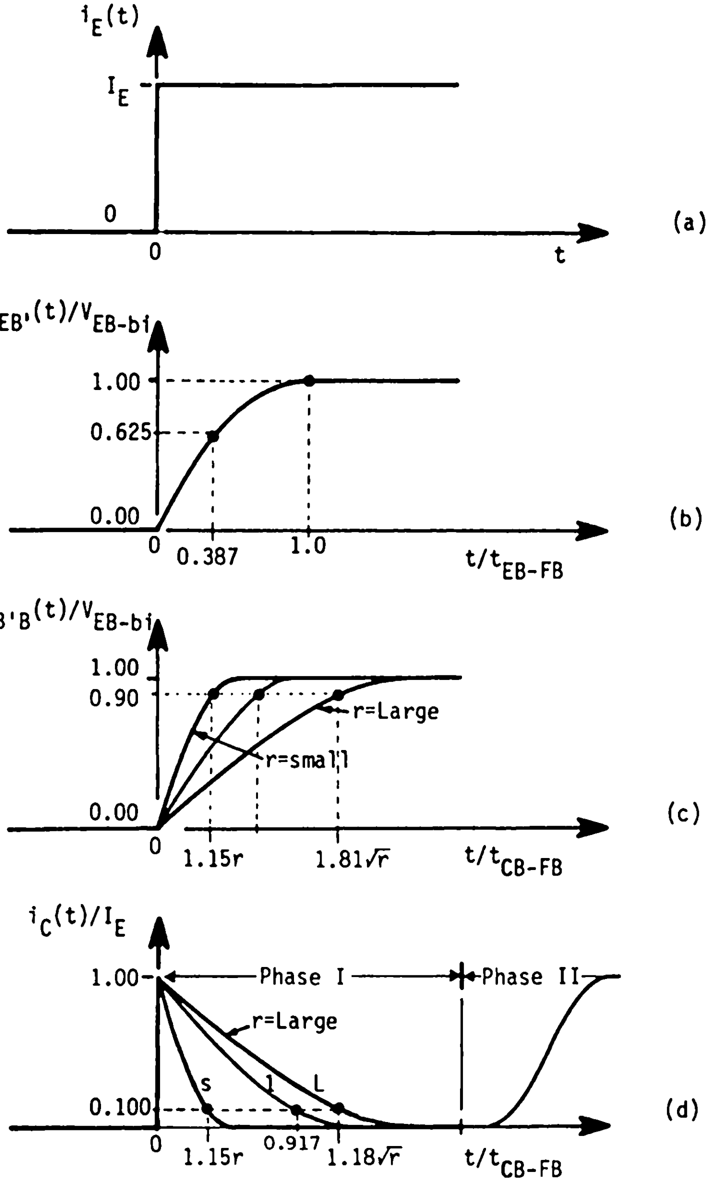


Fig.752.2 Waveforms of the common-base phase I turn-on short-circuit transient in a p/n/p BJT with a step emitter current drive. (a) The input emitter current step. (b) The emitter-base junction space-charge layer voltage, $v_{EB}(t)$. (c) The collector-base junction space-charge layer voltage or the voltage across the base resistance, $v_{B'B}(t)$. (d) The collector current.

The division of the collector current waveform into two phases is particularly sharp if the **minority-carrier** (holes) **base transit time** due to diffusion delay,

$$t_B = X_B^2/2D_B \approx \quad (10\text{ps}) \quad (\text{collector current}) \qquad (752.1)$$

is much larger than the time constant of charging the space-charge layer capacitance of the emitter-base junction by the majority carriers through the base resistance,

$$t_{b'c} = r_{b'}C_{cbt0} \approx \quad (10\text{ps}) \quad (\text{collector current}). \qquad (752.2)$$

C_{cbt0} is the collector-base junction space-charge layer capacitance at zero applied bias. The third time constant is the charging time constant of the emitter-base junction space-charge layer capacitance by the majority-carriers through the base resistance,

$$t_{b'e} = r_{b'}C_{ebt0} \approx \quad (10\text{ps}) \quad (\text{emitter-base voltage}). \qquad (752.3)$$

C_{ebt0} is the emitter-base junction space-charge layer capacitance at zero applied bias. (Numbers given in the parenthesis are typical values using $X_B = 0.1\mu m$, $D_B = 5cm^2/s$ or $\mu_B = 200cm^2/V\text{-}s$, and $N_{BB} = 10^{17}cm^{-3}$ which are to be computed in the ensuing discussions. The second parenthesis indicates its dominance on a current or voltage waveform.)

Two additional characteristic times also affect the initial switching transient. They are the time to **flat-band** the potential barrier of the emitter-base and collector-base junctions by a constant emitter current density of $J_E = I_E/A_E$. They are defined by

$$t_{EB\text{-}FB} = \sqrt{2q\varepsilon_s N_{HE} V_{EB\text{-}bi}}(A_E/I_E) = (qV_{EB\text{-}bi}/kT)r_eC_{eb0} \approx 40\cdot r_eC_{eb0}$$

$$\approx (10\text{ps at } 10^4\text{A/cm}^2) \quad (\text{emitter-base voltage}) \qquad (752.4)$$

and

$$t_{CB\text{-}FB} = \sqrt{2q\varepsilon_s N_{HC} V_{CB\text{-}bi}}(A_C/I_E) \approx (10\text{ps at } 10^4\text{A/cm}^2) \quad (\text{collector current}). \qquad (752.5)$$

Note that the flat-band time constant is much larger than the small-signal time constant, about 40 times larger as estimated in (752.4), because it spans the full built-in voltage while small-signal voltage is only kT/q. In practice, BJTs are seldom driven so hard to flat-band the emitter-base junction (only to about $V_{EB\text{-}bi}/2$). However, these time constants give a useful indication of the initial delay since the emitter-base space-charge layer capacitance must be charged up to a sufficiently high forward voltage before a large number of minority carriers can be injected from the emitter into the base layer.

The sharp division into two phases enables a clear description of the individual processes. Thus, we shall make the assumption of a fairly thick base layer (1μm instead of 0.1μm) so that t_B (100ps) is much larger compared with $r_{b'}\cdot C_{cbt0}$ (10ps) and the two phases are sharply separated.

Phase I CB Turn-on Transient - Charge-Control Solution

During phase I, immediately after an emitter current step is applied, few holes have diffused through the base layer to give the transistor action. Thus, the device can be considered as a T network consisting of two back-to-back connected rectifying diodes in the series arm and a resistance in the shunt arm. The majority-carrier **moving into** the space-charge layer is responsible for charging up the emitter-base junction in the forward direction. This is illustrated by the two arrows of path I pointing towards the emitter-base junction space-charge layer in Fig.752.1(b): one arrow for holes in the p-emitter moving towards the left boundary and the second arrow for electrons in the n-base moving towards the right boundary of the emitter-base junction space-charge layer. The majority carriers **moving away** from the space-charge layer are responsible for charging up the collector-base junction in the reverse direction. This is illustrated by the two arrows of path I in Fig.752.1(b) pointing away from the collector-base junction space charge layer. The collector-base junction voltage eventually reaches the steady-state value determined by the magnitude of the input current and the base resistance, $r_{b'}$. In the following paragraphs, an exact solution is obtained for the voltage and current waveforms of phase I from majority carriers flowing in path I when minority carrier diffusion in path II through the base can be neglected. Exact analytical solution is possible when the voltage dependence of the space-charge layer capacitance of the emitter-base and collector-base junction space-charge layers, C_{ebt} and C_{cbt}, has a simple voltage dependence, such as the parabolic voltage dependence in an abrupt p/n junction that has a spatially constant dopant donor and acceptor impurity concentration profile used as the illustration in chapter 5. After the exact phase I solution is obtained, the analytical solutions will be obtained for the simultaneous presence of paths I and II. Analytical solutions are possible if it is assumed that the capacitances are constant independent of voltage.

Assuming a step-constant impurity profile for both the emitter-base and collector-base junctions and referring to Fig.752.1(b), the charge-control equations (751.9) and (751.10) are then:

$$\begin{aligned}
i_E(t) &= -A_E(d/dt)[qN_{EE}]\cdot X_{EB-e} \\
&= -A_E(d/dt)[qN_{EE}]\cdot[N_{BB}/(N_{EE}+N_{BB})]\sqrt{2\varepsilon_s[V_{EBbi}-v_{EB'}(t)]/qN_{ME}} \\
&= -A_E(d/dt)[\sqrt{2\varepsilon_s qN_{ME}[V_{EB-bi} - v_{EB'}(t)}] \\
&= -t_{EB-FB}(d/dt)\sqrt{1 - [v_{EB'}(t)/V_{EB-bi}}]I_E \qquad (752.6)
\end{aligned}$$

and

$$i_C(t) = +\, t_{CB\text{-}FB}(d/dt)\sqrt{1 - [v_{CB'}(t)/V_{CB\text{-}bi}]}\,I_E(A_C/A_E)$$

$$= +\, t_{CB\text{-}FB}(d/dt)\sqrt{1 + [v_{B'B}(t)/V_{EB\text{-}bi}]}\,I_E(A_C/A_E). \qquad (752.7)$$

These can be solved simultaneously with the circuit equation from Fig.752.1(c)

$$i_E(t) = i_C(t) + v_{B'B}(t)/r_{b'}. \qquad (752.8)$$

The waveform of charging up of the internal emitter-base junction voltage by the applied emitter current step, I_E, can be readily derived by integrating (752.6) which gives

$$v_{EB'}(t)/V_{EB\text{-}bi} = 1 - [1 - (t/t_{EB\text{-}FB})^2]. \qquad (752.9)$$

This simple result is plotted in Fig.752.2(b). It also shows the limit of validity given by $v_{EB'}(t)/V_{EB\text{-}bi} < 0.625$ and $t/t_{EB\text{-}FB} < 0.387$, which is derived next.

The assumption of negligible hole diffusion current through the base layer (path II) will no longer hold when the internal emitter-base junction voltage reaches the injection threshold or the base diffusion-recombination current becomes comparable to the space-charge-layer recombination current and the current to charge up positively the emitter-base transition layer capacitance. Let this condition be $j_b/j_{eb} = 100$. Then, using

$$j_b/j_{eb} = (D_B\tau_{eb}/X_BX_{EB})(n_i/N_{BB})\exp(qV_{EB'}/2kT)$$

and

$$V_{EB\text{-}bi} = (kT/q)\log_e(N_{BB}N_{CC}/n_i^2),$$

we have

$$\frac{V_{EB'}}{V_{EB\text{-}bi}} = \frac{\log_e[10^4(X_BX_{EB}/L_B^2)^2(N_{BB}/n_i)^2]}{\log_e[N_{EE}N_{BB}/n_i^2]} \qquad (752.10)$$

$$= \frac{\log_e[10^4(10^{-4}\text{x}10^{-5}/10^{-5})^2(10^{17}/10^{10})^2]}{\log_e[10^{19}\text{x}10^{17}/10^{10}\text{x}10^{10}]} \qquad (752.11)$$

$$= (\log_e 10^{10})/(\log_e 10^{16}) = 10/16 = 0.625. \qquad (752.11\text{A})$$

The numerical values used in (752.11) are: $X_B = 10^{-4}$cm, $X_{EB} = 10^{-5}$cm = constant, $L_B^2 = D_B\tau_B = 10\text{x}10^{-6}\text{cm}^2$, $N_{BB} = 10^{17}\text{cm}^{-3}$, $n_i = 10^{10}\text{cm}^{-3}$, and $N_{EE} = 10^{19}\text{cm}^{-3}$. The result, (752.11A), is used in the analytical solution of $v_{EB'}(t)$, (752.9), to give the time at which the minority carrier recombination-diffusion current in the base layer is 100 times larger than the recombination current in the emitter-base space-charge layer. Denoting this time as the threshold injection time by t_{TH} as in chapter 5 on p/n junction diode, then $0.625 = 1 - [1 - (t/t_{EB\text{-}FB})^2]$ which gives

$$t = 0.378\cdot t_{EB\text{-}FB} \equiv t_{TH}. \qquad (752.12)$$

Using the numerical values assumed, the flat-band time constant of the emitter-base junction defined by (752.4) is

$$t_{EB-FB} = \sqrt{2q\varepsilon_s N_{ME} V_{EB-bi}}\cdot(A_E/I_E) \qquad (752.4)$$

$$= \sqrt{2\mathrm{x}1.6\mathrm{x}10^{-19}\mathrm{x}10^{17}\mathrm{x}(0.2585)\log_e(10^{19}\cdot 10^{17}/10^{12})}10^{-7}$$

$$= 17.78(\mathrm{nA}\text{-}\mu\mathrm{s})/I_E(\mathrm{nA}). \qquad (752.13)$$

The phase I ending time or phase II starting time is then

$$t_{TH} = 0.378\cdot t_{EB-FB} = 6.72(\mathrm{nA}\text{-}\mu\mathrm{s})/I_E(\mathrm{nA}) \qquad (752.13A)$$

$$= 6.72\mathrm{ps} \quad \text{at } I_E = 1\mathrm{mA} \qquad (752.13B)$$

where a 10 square micron emitter area is assumed with a length of 10μm and width 1μm: $A_E = L_E \cdot W_E = 10^{-3}\cdot 10^{-4}\mathrm{cm}^2 = 10^{-7}\mathrm{cm}^2$. The numerical result here shows that the phase I transient is fast if the input current step is large. For example, if the emitter drive has a current step of $I_E = 1.0\mathrm{mA} = 10\mathrm{kA/cm}^2$, then $t_I = 6.72\mathrm{ps}$. Because the phase I transient passes very quickly, it has been ignored in past and even recent textbooks. However, this can no longer be neglected in present and future ultrahigh speed BJT and bipolar integrated circuits which are switched at a higher current density (than the example above) and higher speed. For example a 0.1μm base BJT gives a base transit time of $t_B = X_B{}^2/2D_B = (10^{-5})^2/2\mathrm{x}5 = 10\mathrm{ps}$. Thus, the emitter-base junction charging or threshold injection time of 1ps has become a significant fraction of the base transit time.

We now compute the waveforms of the internal-base/collector voltage, $v_{B'B}(t)$, due to the presence of the base spreading resistance, and then the collector current, $i_C(t)$. These can be obtained from (752.7) and (752.8). Substituting $i_C(t)$ from (752.7) into (752.8), we have the following normalized differential equation for $v_{B'B}(t)$ which can be solved by elementary integration.

$$i_E(t)/I_E = i_C(t)/I_E + (V_{CB-bi}/r_{b'}I_E)(v_{B'B}/V_{CB-bi}) \qquad (752.14)$$

$$= 1 = (d/dt)\sqrt{1+v(t)} + v(t)/r \qquad (752.14A)$$

where

$$v(t) \equiv v_{B'B}(t)/V_{CB-bi} \qquad (752.15A)$$

$$r \equiv r_{b'}I_E/V_{CB-bi} \equiv t_{b'c}/t_{CB-FB} \qquad (752.15B)$$

$$T \equiv t/t_{CB-FB} \qquad (752.15C)$$

and

$$T' \equiv T/r \equiv t/t_{b'c} = t/(r_{b'}C_{cb0}) \qquad (752.15D)$$

The solutions are

$$T' = T/r = \frac{1}{\sqrt{(1+r)}}\left[\operatorname{arctanh}\frac{\sqrt{1+v(t)}}{\sqrt{(1+r)}} - \operatorname{arctanh}\frac{1}{\sqrt{(1+r)}}\right] \qquad (752.16)$$

$$v(t) = \left[\frac{\sqrt{(1+r)} + (1+r)\tanh[\sqrt{(1+r)}T']}{\sqrt{(1+r)} + \tanh[\sqrt{(1+r)}T']}\right]^2 - 1 \qquad (752.17)$$

$$= r\left[\frac{\tanh[\sqrt{(1+r)}T']}{\sqrt{(1+r)} + \tanh[\sqrt{(1+r)}T']}\right]\left[2 + \frac{r\cdot\tanh[\sqrt{(1+r)}T']}{\sqrt{(1+r)} + \tanh[\sqrt{(1+r)}T']}\right]$$

$$(752.17A)$$

$$i_C(t)/I_E = i_E(t)/I_E - v(t)/r = 1 - v(t)/r \qquad (752.18)$$

$$= 1 - \left[\frac{\tanh[\sqrt{(1+r)}T']}{\sqrt{(1+r)} + \tanh[\sqrt{(1+r)}T']}\right]\left[2 + \frac{r\cdot\tanh[\sqrt{(1+r)}T']}{\sqrt{(1+r)} + \tanh[\sqrt{(1+r)}T']}\right].$$

$$(752.18A)$$

If $r = r_b' I_E/V_{CB-bi} = 0$, the following exact solution is obtained.

$$i_C(t)/I_E = \exp(-2T') = \exp[-2t/(r_b' C_{cb0})]$$

$$(752.18B)$$

The asymptotic result for $r \to 0$ or $(r_b \cdot I_E/V_{CB-bi}) \to 0$ given by (752.18B) is especially interesting. It shows that the decay of the initial output current due to capacitance feedthrough is twice as fast as the initial RC time constant, $r_b \cdot C_{cb0}$. This is not unexpected because the collector-base junction capacitance decreases with increasing reverse junction voltage. However, if the transistor base layer is very thin, then the base spreading resistance would increase with time because the quasi-neutral base layer is thinning as the collector-base junction space-charge layer is thickened by the charging reverse current step. Then, the decreasing capacitance and increasing base resistance would give a constant RC product or time constant. If the emitter is floating and a reverse base-collector voltage step is applied, then the response is truly linear with a time constant twice the initial RC time constant although the system or its elements are nonlinear.

The time constant for phase I may be defined as the time required for the collector current to drop to 10% of I_E or $i_C(t=t_I)/I_E = 0.1$. Plug this into (752.18),

$$v(t_I)/r = 1 - i_C(t_I)/I_E = 1.0 - 0.1 = 0.9 \qquad (752.19)$$

which is then used in (752.16) to give the phase I duration:

$$t_I/t_{b'c} = t_I/(r_{b'}C_{cb0}) = T_{I'} = T_I/r$$

$$= \frac{1}{\sqrt{(1+r)}}\left[\text{arctanh}\frac{\sqrt{1+0.9r}}{\sqrt{(1+r)}} - \text{arctanh}\frac{1}{\sqrt{(1+r)}}\right] \qquad (752.20)$$

$$\simeq \text{arctanh}(0.45/0.55) = 1.151 \qquad r \ll 1 \qquad (752.21A)$$

$$= \qquad = 0.917 \qquad r = 1 \qquad (752.21B)$$

$$\simeq [\text{artanh}\sqrt{0.9}]/\sqrt{r} = 1.182/\sqrt{r} \qquad r \gg 1\ . \qquad (752.21C)$$

These show that t_I is nearly equal to $r_{b'}C_{cb0}$ (the charging time constant of the zero-bias collector capacitance C_{cb0} through the base resistance $r_{b'}$) in the practical range of $0 < r=(r_{b'}I_E/V_{CB\text{-}bi}) < 1$. This is twice shorter than a truly exponential decay from a voltage independent collector-base space-charge layer capacitance which has a normalized time constant of $T_{I'} = \log_e 10 = 2.303$ or exactly twice longer than the above realistic case.

Using the numerical values just assumed, the base series conductance or the reciprocal base series resistance is

$$g_{b'} = q\mu_p P_B(X_B L_E/W_E/2) = 1/r_{b'} =$$

$$= 1.6\text{x}10^{-19}\text{x}250\text{x}10^{17}\text{x}(10^{-5}\text{x}10^{-3}/10^{-4}/2) = 8\text{x}10^{-4}\text{S} = 1/(1250\Omega)$$

and the collector-base junction capacitance at $v_{CB'}=0$ and $t=0$ is

$$C_{cb0} = \sqrt{2\varepsilon_s q N_{MC}/V_{CB\text{-}bi}}A_C$$

$$= \sqrt{2\text{x}11.7\text{x}8.854\text{x}10^{-14}\text{x}1.6\text{x}10^{-19}\text{x}10_{16}/0.0259\text{x}\log_e(10^{17}\text{x}10^{16}/10^{20})}A_C$$

$$= 6.545\text{x}10^{-7}(\text{F/cm}^2)A_C = 6.545\text{x}10^{-7}(10\text{x}10^{-4}\text{x}2\text{x}10^{-4}) = 13.09\text{fF}$$

where we have assumed that the collector is twice as wide as the base to allow contact areas. Then, the time constant of charging the collector-base capacitance through the base resistance, defined in (752.3), is

$$t_{b'c} = r_{b'}C_{cb0} = 1250\text{x}13.09\text{fs} = 16.36\text{ps}. \qquad (752.22)$$

This sample result again shows that phase I lasts only a few tens picoseconds. But it becomes important when the base transport time, $t_B=X_B^2/2D_B$, is reduced to the picosecond range which is occurring in present and future generations of ultrahigh speed BJT switches for mainframe computer, three-dimensional graphics processor, and other very high speed applications.

Phase II (including Phase I) CB Turn-on Transient - Lump Model

The phase II turn-on transient is due to the minority carrier (holes) diffusion delay during transit through the base layer, X_B, whose transit time is given by (752.1). These minority carriers discharge the collector-base junction capacitance which was charged up to a voltage of $r_{b'} \cdot I_E$ during phase I. The exact charge-control solution requires solving a nonlinear differential equation of the form $dz/dt + z^2 = A + B\exp(-\omega t)$ which we shall not pursue. Instead, we shall obtain the analytical solution of the waveforms valid for both phases I and II, which exists if the collector-base junction capacitance is assumed to be a constant, independent of voltage.

The physical circuit model was given in Fig.752.1(b) where the diffusion T-section in the base for phase II is replaced by a dependent current source, $\alpha_f i_E(t)$ in Fig.752.3. The waveform was shown in Fig.752.2(d). It was expanded in Fig.752.4(a)-(d) where (a) shows the input current, (b) shows the emitter-base junction voltage, (c) shows the internal-base/collector voltage, $v_{B'B}(t)$ across the base spreading resistance $r_{b'}$, and (d) shows $i_C(t)$.

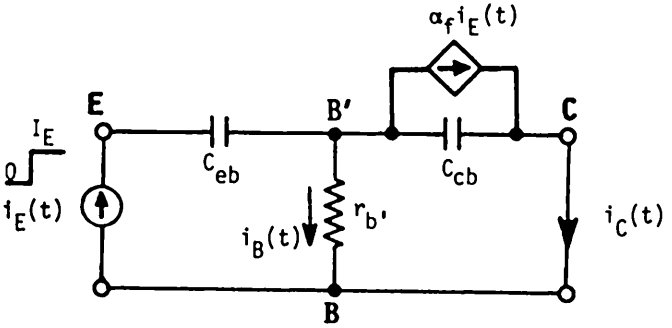


Fig.752.3 The common-base large-signal Tee equivalent circuit for analyzing the short-circuit transient of a BJT drive by an emitter current step where the two capacitances represent the nonlinear charge-voltage relationships of the two p/n junctions.

The node equation at the internal base node B' is

$$I_e = V_{b'b}/r_{b'} + \alpha_f I_e + sC_c V_{b'b} \qquad (752.23)$$

where s is the Laplace transform variable. Assume a step emitter current, $I_e = I_E/s$, and a 1-pole approximation to the base transport factor, $\alpha_f = \alpha_{fo}/(1+s/\omega_B)$, then (752.23) becomes

$$V_{b'b} = (I_E/C_c)(1/s)[1/(s+\omega_{b'c})][1 - \alpha_{fo}\omega_B/(s+\omega_B)]. \qquad (752.24)$$

The solutions are

$$\begin{aligned} v_{B'B}(t)/(r_{b'}I_E) = 1 - \alpha_{fo} & \\ & - [1 + \alpha_{fo}\omega_B/(\omega_{b'c}-\omega_B)]\exp(-\omega_{b'c}t) \\ & + [\alpha_{fo}\omega_{b'c}/(\omega_{b'c}-\omega_B)]\exp(-\omega_B t) \end{aligned} \qquad (752.25)$$

and

$$i_C(t)/I_E = 1 - v_{B'B}(t)/(r_{b'}I_E). \qquad (752.26)$$

These solutions of the linearized system are only rough approximations to the exact solutions. We can examine the errors by considering the ideal case of $r_{b'}\to 0$ or $r_{b'}C_c << t_B = \omega_B^{-1}$ because the exact phase I solution was obtained in (752.18B). For short times during phase I, $\omega_{b'c}^{-1} \approx t << \omega_B^{-1}$ and for long times during phase II, $\omega_{b'c}^{-1} << t \approx \omega_B^{-1}$; then (752.26) and (752.25) give respectively

$$\text{(I)}\ i_C(t)/I_E \approx 1 - \exp(-\omega_{b'c}t) \qquad 0 \le t \approx r_{b'}C_c << \omega_B^{-1} \qquad (752.26\text{A})$$

and

$$\text{(II)}\ i_C(t)/I_E = \alpha_{fo}[1 - \exp(-\omega_B t)] \qquad r_{b'}C_c << t \approx \omega_B^{-1} \qquad (752.26\text{B})$$

They are the expected one-pole lump circuit solutions, but they are in serious error. For example, the lump circuit solution for phase I given by (752.26A) is in grave error compared with the exact solution, (752.18B). The time constant is exactly 2 times too large because the lump circuit had assumed a voltage-independent constant collector-base junction capacitance.

The lump circuit solution for phase II (752.26B) is also in serious error as indicated by a comparison with the exact solution plotted in Figs.752.4(c) and (d). The exact analytical solutions for phase II will be described in later paragraphs. In figures (c) and (d), we note that the exact solution of the collector current predicts an initial delay of $t_d = 0.26t_B = 0.26\omega_B^{-1} = 0.26\cdot(2D_B/X_B^2)$ to reach 10% of the final value due to minority carrier diffusion delay through the base layer. However, the lump solution just derived, (752.26B), predicts a delay of $t_d = -\log_e(0.9)t_B = 0.105t_B$ which is 2.5 times too small. The basic physics-cause of the error is the one-pole, one-lump, or one-time constant approximation, just described in the foregoing analysis that gave (752.26B), which can only account for a gradual spatial variation of the hole distribution in the base layer. It cannot account for the very sharp spatial variation of the concentration of holes which are all concentrated near the emitter edge of the quasi-neutral base layer during the initial period of phase II. For this reason, the lump solution gives a better approximation at long times than at short times. For example, the exact solution gives a rise time to 90% of the final value of $2.0622t_B$ while the lump solution, (752.26B), gives $(\log_e 10)t_B$ or $2.303t_B$ which is in error by only 11.16% instead of $0.26/0.105=250\%$.

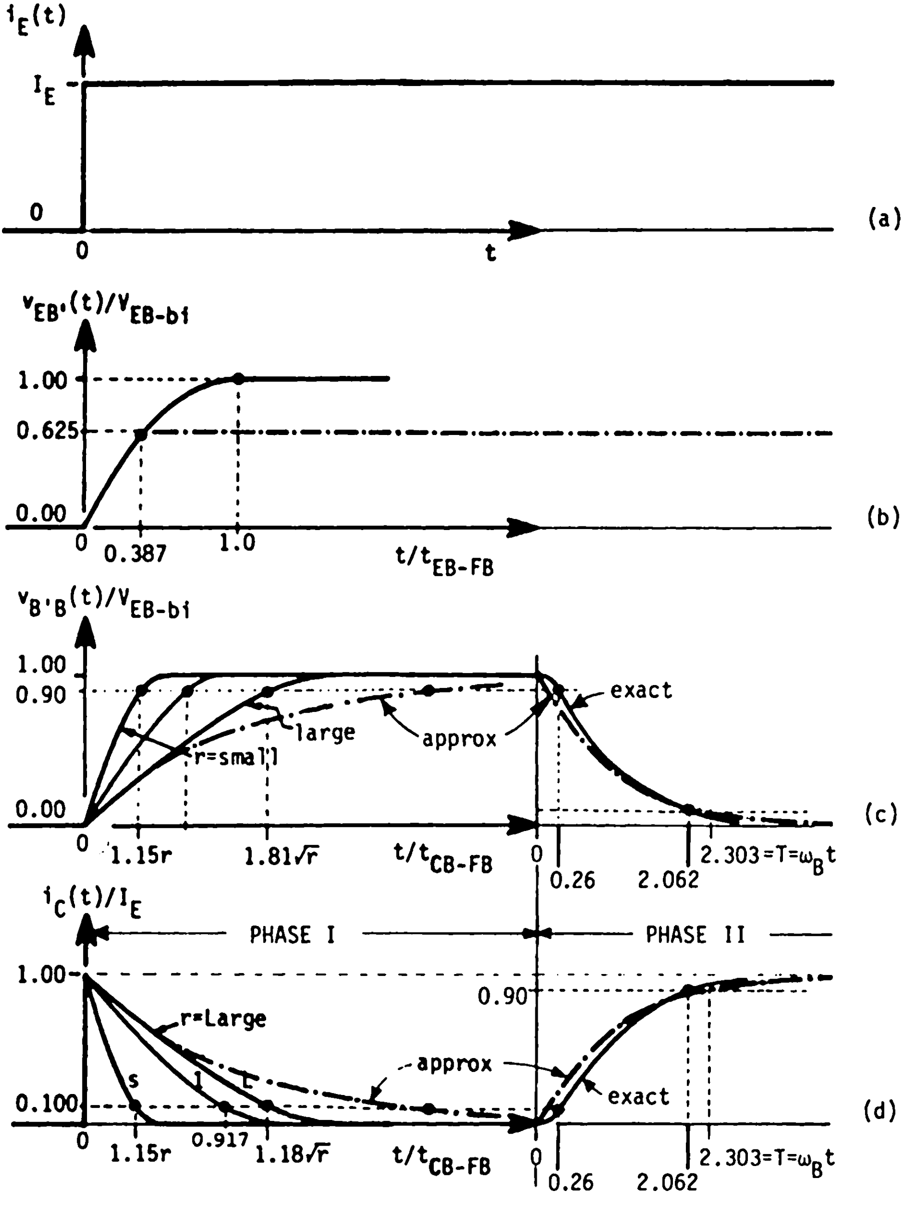


Fig.752.4 The waveforms of the lump approximation and exact solutions of the phase I and II CB turn-on transients in a BJT driven by a step emitter current. (a) Step emitter current drive. (b) Exact phase I emitter-base voltage. (c) Internal base-collector voltage. (d) Collector current.

After an outline description of the exact solutions to be given immediately in the following paragraphs, we shall describe a novel charge-control technique that gives much better solutions than the one lump solution just described. This improved method was invented based on simple basic device physics that anticipated the time dependence of the spatial distribution of minority carriers (holes) in the base during the transient.

Phase II CB Turn-On Transient - Exact Solutions

The exact waveforms given in Figs.752.4(c) and (d) are obtained as follows. They are from the exact analytical solutions which are obtained by neglecting all the parasitics so that the only mechanism during phase II is the minority carrier diffusion delay in the base layer. In this case, the minority carrier (holes) transport in the base layer is governed by two minority carrier transport equations: the diffusion current and continuity equations. From the Shockley equations (350.2) and (350.4), the one-dimensional equations are

$$j_P(x,t) = -qD_p\partial p(x,t)/\partial x \qquad (752.27)$$

and

$$q\partial p(x,t)/\partial t = -\partial j_P(x,t)/\partial x + q(g_P - r_P) \qquad (752.28)$$

which can be combined by eliminating j_P to give the diffusion equation (751.3B),

$$q\partial p(x,t)/\partial t = +qD_p\partial^2 p(x,t)/\partial x^2 + q(g_P - r_P). \qquad (752.29)$$

For the common-base short-circuit turn-on transient with a constant emitter current step, the boundary, initial and final steady-state conditions are as follows. The constant emitter current requires

$$j_P(x=0,t) = J_E \equiv J_P = -qD_p\partial p(x=0,t)/\partial x. \qquad (752.30A)$$

The short-circuit collector-base junction requires

$$p(x=X_B,t) - P_B = 0. \qquad (752.30B)$$

The zero initial emitter-base bias or current requires

$$p(x,t<0) - P_B = 0. \qquad (752.30C)$$

The initial collector current must be zero due to the diffusion delay through the base,

$$j_P(x=X_B,t=0) = 0. \qquad (752.30D)$$

The final collector current must be equal to that of the emitter current less any recombination loss in the base. If base recombination is neglected, which is a good approximation for high-beta and high-frequency transistors, then

$$j_P(x=x_B,t\to\infty) = j_P(x=0,t\to\infty) = J_E \equiv J_P. \qquad (752.30E)$$

Finally, if recombination and generation are neglected, then the terms $q(g_P - r_P)$ in (752.29) can be dropped and it simplifies to

$$\partial p(x,t)/\partial t = + D_p \partial^2 p(x,t)/\partial x^2. \qquad (752.30F)$$

In some practical problems, the generation term is quite important, for example, in solar cell and photonic devices. The generation term must also be retained in high electric fields when electrons and holes are generated via interband impact generation of electron-hole pairs by the energetic electrons. Recombination must be retained in the common-emitter circuits to be described later.

Two series forms of the exact solution for the collector current can be obtained, one with good convergence at short times using Laplace transform technique, and the second, at long times using the separation of variable technique. We shall provide an abbreviated description of the algebraic procedure which is quite straightforward. The fast-convergent short-time solution is obtained by (1) making a Laplace transform of the second order partial differential equation for hole diffusion (752.30F), from the time domain to complex frequency domain (indicated by s), $p(x,t) \to p(x,s)$, converting it to a second-order ordinary differential equation, (2) finding the general space solution of $p(x,s)$ that contains three terms and two coefficients from the order (second) and the known initial hole distribution $p(x,t=0^+)$, (3) determining the constants by the boundary and initial conditions, (4) substitute the solution of the two constants into the general solution in (2) to give the formulae of $p(x,s)$ and $j_P(x,s)$, and (5) writing down the answer in the time domain, $p(x,t)$ and $j_P(x,t)$, by taking the inverse Laplace transform of $p(x,s)$ and $j_P(x,s)$ using a Laplace transform table. The five steps are listed in the following equations (752.3n) where $n=1$ to 5.

$$sp(x,s) - p(x,t=0^+) = D_p \partial^2 p(x,s)/\partial x^2 \qquad (752.31)$$

$$p(x,s) = p(x,t=0+)/s + A\exp[+\sqrt{(D_p/s)}x] + B\exp[-\sqrt{(D_p/s)}x] \qquad (752.32)$$

$$p(x,t=0+) = 0 \qquad (752.33A)$$

$$p(x=X_B,s) = A\exp[+\sqrt{(D_p/s)}X_B] + B\exp[-\sqrt{(D_p/s)}X_B] = 0 \qquad (752.33B)$$

$$J_P = - qD_p \partial p(x,t)/\partial x \Big|_{x=0} = - qD_p\sqrt{D_p/s}(A - B) \qquad (752.33C)$$

$$p(x,s) = \frac{(J_F/qD_B)\cdot\sinh[\sqrt{(s/D_p)}(X_B-x)]}{s\cdot[\sqrt{(s/D_p)}]\cdot\cosh[\sqrt{(s/D_p)}(X_B)]} \qquad (752.34A)$$

$$j_P(x,s) = \frac{J_F\cdot\cosh[\sqrt{(s/D_p)}(X_B-x)}{s\cdot\cosh[\sqrt{(s/D_p)}(X_B)]} \qquad (752.34B)$$

$$= J_F \sum_{n=0} \frac{(-1)^n}{s} \left[+ \exp\{-\sqrt{(s/D_p)}[(2(n+1)X_B-x]\} + \exp\{-\sqrt{(s/D_p)}[2nX_B+x]\} \right] \qquad (752.34C)$$

$$j_P(X_B,s) = 2J_F \sum_{n=0}^{\infty} \frac{(-1)^n}{s} \exp\{-\sqrt{(s/D_p)}[(2n+1)X_B]\} \qquad (752.34D)$$

$$j_C(t) = j_P(X_B,t)$$

$$= 2J_F\cdot[\mathrm{erfc}(1/\sqrt{2T}) - \mathrm{erfc}(3/\sqrt{2T}) + \mathrm{erfc}(5/\sqrt{2T})] \qquad (752.35)$$

$$\approx 2J_F\cdot\mathrm{erfc}(1/\sqrt{2T}) \qquad (752.35A)$$

where (752.35A) is valid in the short-time range of

$$T = \omega_B t = (2D_B/X_B^2)t < 2.1 \qquad (752.35B)$$

or

$$t < 1.05(X_B^2/D_B). \qquad (752.35C)$$

The short-time solution, (752.35A), can be used to determine the initial delay for the collector current to reach 10% of its final value after a J_E emitter current step is applied. This delay is defined by setting $j_C(t_{0.1}) = 0.1J_F = 2J_F\cdot\mathrm{erfc}(1/\sqrt{2}T_{0.1})$ which gives

$$T_1 = \omega_B t_1 = 2D_B t_1/X_B^2 = 0.260 \text{ (exact)} \qquad (752.36A)$$

or

$$t_1 = 0.13\cdot(X_B^2/D_B) \qquad \text{(exact)} \qquad (752.36B)$$

$$= 0.26\cdot(X_B^2/2D_B) \qquad (752.36C)$$

$$= 0.32\cdot(X_B^2/2.43D_B) \qquad (752.36D)$$

$$t_1 = 0.11\cdot(X_B^2/2D_B) \qquad \text{(1-lump model)} \qquad (752.36E)$$

The lump model with the expression $1 - \exp(-T)$, (752.26A), for the collector current gives an initial delay of $T_1 = 0.105 \approx 0.11$ which substantially underestimates the exact delay of 0.260.

The long-time solution is obtained by the separation of variable technique which assumes that $p(x,t)=p(x)p(t)$ and then expands $p(x)$ in Fourier series to satisfy the boundary and initial conditions. The solution is given by

$$j_C(t)/J_P = 1 - (4/\pi)\{ + \exp[-(\pi^2/8)T] - 3^{-1}\exp[-9(\pi^2/8)T]$$

$$+ 5^{-1}\exp[-25(\pi^2/8)T] - \ldots.\} \qquad (752.37)$$

$$\approx 1 - (4/\pi)\cdot\exp[-(\pi^2/8)T]. \qquad (752.37A)$$

The approximate solution (752.37A) is valid for times longer than $T > 0.15$. Thus, it is an excellent approximation even for fairly short times and is in fact sufficiently accurate to predict the initial delay time which occurs at $T_1 = 0.26$.

The rise time at which the collector current reaches 90% of its final value is obtained from (752.37A) and given by

$$T_{0.9} = \omega_B t_{0.9} = (2D_B/X_B^2)t_{0.9} = 2.062 \qquad (752.38A)$$

or

$$t_{0.9} = 1.031\cdot(X_B^2/D_B) \qquad \text{(exact)} \qquad (752.38B)$$

$$= 2.062\cdot(X_B^2/2D_B) \qquad (752.38C)$$

$$= 2.505\cdot(X_B^2/2.43D_B) \qquad (752.38D)$$

$$t_{0.9} = 2.303\cdot(X_B^2/2D_B) \qquad \text{(1-lump model)} \qquad (752.38E)$$

Equation (752.38E) shows that the total turn-on transient given by the one-lump approximation is $2.303/2.062 = 1.1168$ times or 11.68% too large.

The exact solutions given by (752.35) and (752.37) are computed and plotted in Fig.752.4(d). This figure also gives the one-lump solution, $1-\exp(-\omega_B t)$.

Phase II CB Turn-On Transient - Effects of Parasitics

The linear circuit model shown in Fig.752.5(a) and its approximations during phases I and II in Figs.752.5(b), (c), and (d) are helpful to visualize the effect of the parasitics which will be discussed in the following paragraphs.

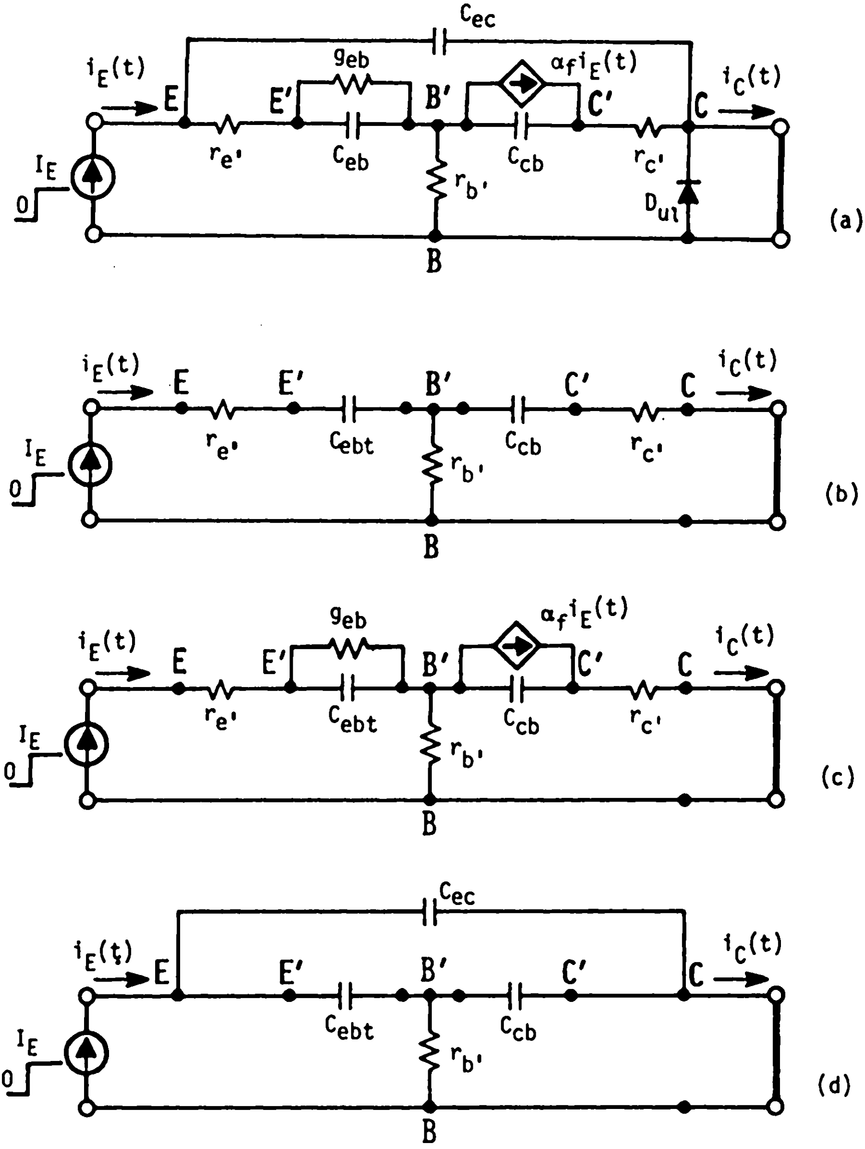

C_{ec}
$i_E(t)$
E
g_{eb}
E'
B'
$\alpha_f i_E(t)$
C'
C
$i_C(t)$
I_E
0
$r_{e'}$
C_{eb}
$r_{b'}$
C_{cb}
$r_{c'}$
D_{ul}
B
(a)
C_{ebt}
(b)
(c)
(d)

Fig.752.5 The effects of parasitic contact resistances, stray feedthrough capacitances and underlap collector-base diode on the CB short-circuit turn-on transient. (a) The complete circuit model. Effects of series resistances on (b) phase I neglecting phase II and (c) phase II neglecting I. (d) Effects of feed-through capacitance on phase I neglecting phase II.

The emitter series resistance obviously has no effect since it is in series with a constant current source. The collector series resistance has no effect on the transmission of the initial current step because the emitter drive is a current source. However, it will prolong the decay of the phase-I collector current by the discharging time constant, $r_{c'}C_{cbt}C_{ebt}/(C_{cbt}+C_{ebt})$, as indicated by Fig.752.5(b). It has little effect on the phase-II collector current rise since during this phase, the collector current is driven by a diffusion-delayed dependent-current-source, $\alpha_{fo}/(1+st_B)$, as indicated in Fig.752.5(c).

The collector-base underlap diode and its junction capacitance obviously have no effect since it is short-circuited by the short-circuit load.

Illustrated in Fig.752.5(d), when there is a parasitic coupling or feedthrough capacitance between the collector and emitter terminals, C_{ec}, the collector current has two components: (i) the current flowing through the series emitter-base and collector-base junction space-charge capacitances just discussed, and (ii) the current flowing through the parasitic coupling capacitance. C_{ec} will affect only the phase I collector current decay and not phase II until phase II becomes so short that the collector current transient cannot be separated into two distinguishable phases.

During phase I, this parasitic feedthrough capacitance, C_{ec}, will not affect the initial rise of the collector current which faithfully reproduces the emitter current step. However, it does not have a resistance to divide or leak off the emitter drive current like $r_{b'}$ which diverts part of the current flowing through the emitter-base space-charge or transition layer capacitance, C_{ebt}. Thus, the collector current from C_{ec} is given by $C_{ec}(dv_{EB}/dt) = C_{ec}(dv_{EB'}/dt) + v_{B'B}$ which will decay to zero when $dv_{EB'}/dt=0$ or when the emitter-base junction is charged up to its final steady-state forward voltage via the base resistance, $r_{b'}$.

Common-Base Turn-Off Transients in BJT

There are four unforced or natural schemes to turn-off the transistor by either open-circuiting or short-circuiting the emitter-base junction or collector-base junction to remove the emitter current drive. Since the collector-base junction was shorted during the steady-state, the simplest switch-off scheme would be to leave the collector-base junction shorted $v_{CB}(t)=0$. It is obvious that of the two remaining turn-off schemes with $v_{CB}(t)=0$, the emitter-base short-circuit scheme will have a faster turn-off transient than the emitter open-circuit scheme. Further speed up can be attained by forcing an emitter current in the reverse direction to draw out the minority carriers stored in the base layer, known as forced turn-off.

The turn-off transients are similarly controlled by the two time constants as the turn-on transients. To turn the BJT off, the minority carriers stored in the base layer must be extracted. This extraction is delayed by diffusion through the base layer towards the collector-base junction space-charge layer in the open-emitter turn-off scheme, and towards both the collector-base and emitter-base junction space-charge layers in the short-circuited emitter-base turn-off scheme. There are two additional transients in addition to the base diffusion delay. One delay is due to the discharge of the collector-base junction space-charge layer capacitor, C_{cbt}, which is rather short since its steady-state voltage or stored charge is very small because it comes from the small d.c. base current, $I_B=(1-\alpha_{fo})I_E\propto I_E/\beta_F$, that flows through the base resistance. The second additional transient arises from discharging the emitter-base junction space-charge layer capacitance, C_{ebt}, which is somewhat longer because it was charged up to a high forward steady-state voltage by I_E. The results of both the charge-control and exact solutions of the open-emitter and shorted emitter-base turn-off transient are described in the following paragraphs.

Charge-Control Solutions of CB Turn-Off Transients

As just stated, the turn-off transients in collector current and emitter or base voltage originate from the process of drawing out the stored holes in the base layer through the collector terminal, and also the emitter terminal (shorted emitter-base only). The conventional 1-lump circuit model cannot predict the turn-off transients accurately when an abrupt change occurs to initiate the turn-off, for example, short-circuiting the emitter-base terminals which abruptly changes the terminal voltage to zero, and open-circuiting the emitter which abruptly changes the emitter current to zero. The error is especially large at and near the beginning of the turn-off transient because the reduction of the minority carrier (holes) concentration is localized at the emitter junction $x=0$ where the terminal current or voltage change occurs. The drawn out holes are very localized and hence causes very abrupt spatial changes. For the open emitter turn-off case, this is illustrated by the minority carrier (holes) distribution in the base layer shown in Fig.752.6(a). The initial drawn out holes are in the cross-hatch shaded right triangle enclosed by the $t=0$, $t=1$ and $x=0$ lines. The charge-control solution gives a better approximation at longer times since holes are drawn out from the entire base layer more uniformly, such as the shaded and slanted triangle shown in Fig.752.6(a) whose edges are the $t=0$, $t=2$ and $x=0$ lines for the open-emitter turn-off case. The magnitude of the error will be illustrated by comparing the charge-control solutions with the exact solutions of the open and shorted emitter cases to be described in the following paragraphs.

First, note a general result regardless of model and geometry details: the open-emitter turn-off transient is exactly the complement of the turn-on transients just obtained. Thus, from the turn-on charge-control solution, (752.26B), the **open-emitter turn-off transient** is given by the one-time constant solution

$$i_C(t) = \alpha_{fo}I_E\exp(-\omega_B t). \tag{752.39}$$

The direct derivation of this result will now be described to verify that the complementary solution is correct.

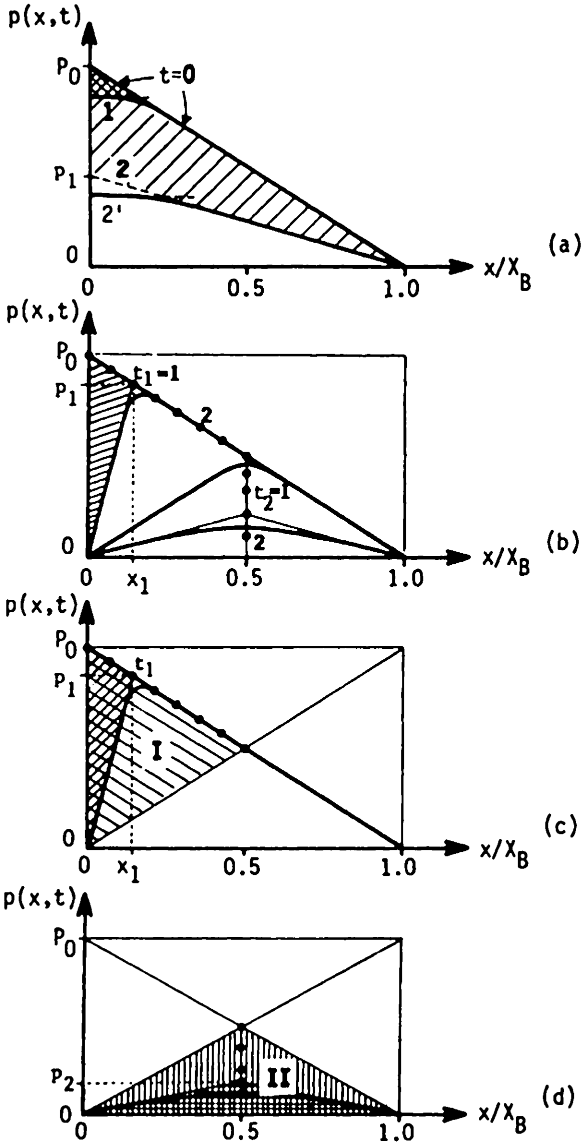


Fig.752.6 Minority carrier (holes) distribution in the base layer for the charge-control analysis of CB turn-off transients. (a) Open emitter. Shorted emitter giving: (b) both phase I and II, (c) phase I only, and (d) phase II only.

From the longer time triangle enclosed by the $x=0$, $t=0$ and $t=2$ lines shown in Fig.752.6(a), the charge-control equations are given below.

$$J_P(x=0,t) - J_P(x=X_B,t) = \int_0^{X_B} q[dp(t)/dt]dx = dq_P/dt$$

$$= \quad 0 \quad - \quad qD_pp_1/X_B = (d/dt)[q(P_0 - p_1)X_B/2]. \qquad (752.40)$$

Normalizing this equation to the initial steady-state emitter current before the emitter is opened, $J_E=qD_pP_0/X_B$, the solution given by (752.39) is immediately obtained. This should not be a bad approximation for long times or phase II since the neglected hole density [depicted by the unshaded small triangle enclosed by $x=0,t=2,t=2'$ in Fig.752.6(a)] is much smaller than the total hole density (the main triangle enclosed by $x=0,t=0,t=2$). The horizontal line of the lower edge of the triangle is due to the boundary condition $j_E(t) = 0 = j_P(x=0,t) = qD_p \cdot dp(x=0,t)/dx$ from the open emitter condition. The goodness of this charge-control approximation for longer times can also be verified by comparing the approximate solution, $\exp(-T)$, with the first term of the exact long-time solution, $(4/\pi)\exp[-(\pi^2/8)T]$, given by the complement of the exact long-time turn-on transient, (752.37A). The approximate time constant given in (752.39) is $8/\pi^2 = 1.2337$ times larger than the exact solution.

The short-time hole distribution at the beginning of the transient, shown by the small cross-hatched triangle ($x=0,t=0,t=1$) at the top in Fig.752.6(a), clearly indicates that the charge-control or one-pole solution is not a good approximation because these holes have to diffuse through the entire base to reach the collector junction at $x=X_B$ in order to reduce the collector current. The charge-control model is inadequate to compute this diffusion delay because it cannot predict the response to a distant source. In this case, the depleted hole charge at $x=0$ is the source which gives the current decrease at the distant collector located at $x=X_B$. The difficulty is further illustrated by an expansion of the short-time exact solution, $i_C(t)/I_E \approx 1 - (2/\sqrt{\pi})\sqrt{(2T)} \cdot \exp(-1/2T)$, which is the complement of short-time turn-on solution given by (752.35A). This complicated dependence on T or $\omega_B t$ confirms the difficulty of its derivation from charge-control analysis using a simple geometrical construction based on one or more simple physical assumptions.

The charge-control analysis of the **short-circuit emitter turn-off** transient can be divided into two phases. These are illustrated by the time dependence of the minority carrier distribution in the base shown in Fig.752.6(b). The loci of the point (p_1,t_1) is indicated in Fig.752.6(c) during phase I and Fig.752.6(d) during phase II. Thus, this two-phase charge-control analysis assumes that there is no drop of the collector current during phase I, and the collector and emitter currents are equal during phase II. This is a good approximation which gives nearly the correct time constants for both phase I and II. The charge control equations can be readily written by inspection of Figs.752.6(c) and (d). The solutions are

$$i_C(t)/I_E = 1 \qquad \text{(Phase I; } T \leq T_1 = 0.193) \qquad (752.41A)$$

$$i_E(t)/I_E = (1/\sqrt{\omega_B t}) - 1 \qquad \text{(Phase I)} \qquad (752.41B)$$

$$i_C(t)/I_E = \exp[-(8D_B/X_B{}^2)(t-t_1)] \qquad \text{(Phase II; } T > T_1 = 0.193) \qquad (752.42A)$$

$$= i_E(t)/I_E. \qquad (752.42B)$$

The initial spike in the emitter current, as indicated by the $1/\sqrt{t}$ dependence in (752.41B), is not unexpected when discharging a capacitance by a short circuit. The inverse square-root time dependence, $t^{-1/2}$ from $1/\sqrt{(\omega_B t)}$, agrees exactly with the expansion of the exact short-time solution, $\sqrt{(2/\pi T)}$. The only error is in amplitude. The charge-control solution is $\sqrt{(\pi/2)} = 1.2533$ or 25.33% too large.

For phase II, the model shown in Fig.752.6(d) is the only choice that will give a time constant, $X_B{}^2/(8D_B)$, which is nearly correct compared with the exact asymptotic solution, $X^2{}_B/(\pi^2 D_B)$, to be derived in the following paragraphs. Other choices of the charge-control model and the lump circuit model would give a time constant from $X_B{}^2/(4D_B)$ to $X_{B2}/(2D_B)$, which is two to four times larger than the exact asymptotic result, $X_2/(\pi^2 D_B)$.

Exact Solutions of CB Turn-off Transients

The exact solutions of these cases can be obtained by the method of separation of variables for short times and by the Laplace transform method for long times. The procedure and arithmetic are identical to those used for the turn-on transients whose solutions were given in (752.35) and (752.37). As stated previously, the open-emitter turn-off transient is exactly the complement of the turn-on transient given by (752.35) and (752.37). Thus, only the short-circuited emitter transients need to be worked out. The results are listed below.

Short-Time Exact Turn-Off Solution of Short-Circuited Emitter-Base

$$i_C(t)/I_E = 1 - (2/\sqrt{\pi})\cdot[2/\sqrt{(2T)}]\cdot\{+\exp[-1/(2T)] + \exp[-9/(2T)] + \exp[-25/(2T)] + \ldots\} \qquad (752.43A)$$

Long-Time Exact Turn-Off Solution of Short-Circuited Emitter-Base

$$i_C(t)/I_E = 2\cdot\{+\exp[-(\pi^2/2)T] - \exp[-4(\pi^2/2)T] + \exp[-9(\pi^2/2)T] - \ldots\} \qquad (752.43B)$$

The exact and approximate charge-control turn-off solutions are plotted in Fig.752.7 for open-emitter and in Fig.752.8 for shorted-emitter.

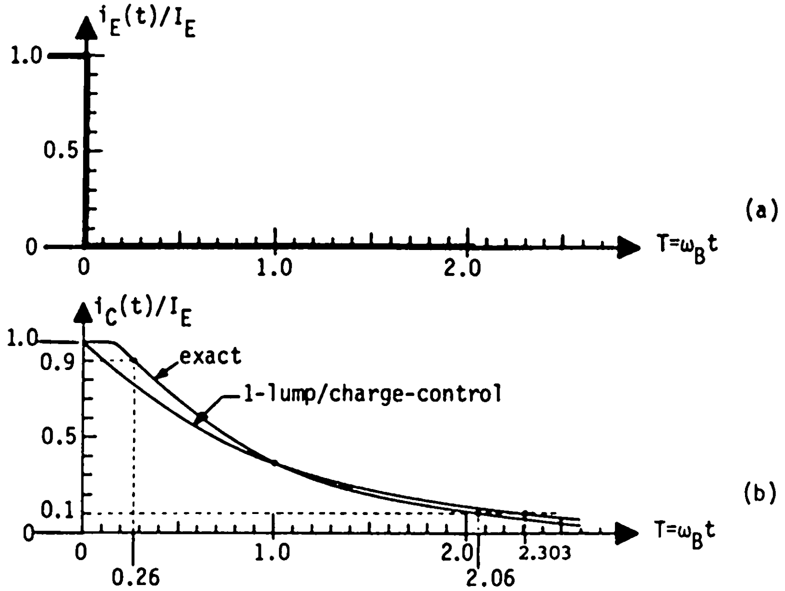


Fig.752.7 Waveforms of the common-base open-emitter shorted-collector turn-off transient in a BJT. (a) The emitter drive current. (b) The collector output current.

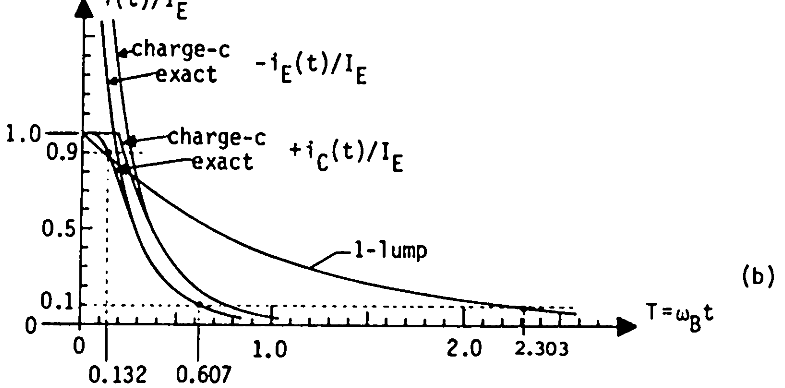


Fig.752.8 Waveforms of the common-base shorted-emitter and shorted-collector turn-off transient in a BJT. (a) The emitter voltage drive. (b) The collector output current.

753 Common-Emitter Large-Signal BJT Switching Transients

In switching circuits, BJTs are often connected in the common-emitter (CE) configuration because this configuration gives both current and voltage gain, just like the small-signal applications. The gain allows a CE BJT inverter to drive more than one CE BJT load (fan-out greater than 1) in order to expand into many branches and strings of cascaded inverters. Multiple fan-out is needed to implement the desired logic functions in a digital logic network. The current gain comes from $\beta_F > 1$, i.e., the collector output current (driving the next stage) is much larger than the base input current (loading the previous stage). Voltage gain comes from both the current gain and the larger collector-emitter output resistance than the base-emitter input resistance. The common-base configuration would not give current gain and hence can drive only one BJT load with diminishing drive capability after several serial or cascaded stages. In this section, we shall analyze the basic limitations on switching speed by considering the ideal case of no load resistance. In the next sections, 76n, switching transients in practical circuits are described which have source, load and internal resistances. These resistances set an upper limit on the collector current in the saturation region of the I_C-V_{CE} characteristics of Fig.732.3(a) if the BJT is over-driven by a large base current. Practical circuits will also be described which were invented to limit large voltage swing and prevent excursion deep into the saturation range.

The charge-control method is used to obtain the transient equations and the terminal current-voltage waveforms of BJT in the ideal CE configuration. The analytical charge-control formulae are then compared with the exact series solutions from Laplace transform and separation of variables methods.

Common-Emitter Turn-On Transient

Consider first a BJT in the CE configuration being turned on by a base current step. The circuit diagram is shown in Fig.753.1(a) with no input and output resistances. The quasi-static low-speed I_C-V_{CE} locus during the switching transient, shown in Fig.753.1(b), is a vertical line since the resistances are zero. The base and collector current waveforms are shown in Fig.753.1(c). Phase I delay [$t<0$ in figure (c)] due to charging the emitter-base junction space-charge layer capacitance is the same as that of the CB configuration. A small reverse (negative) collector current appears while charging the collector-base junction space-charge layer capacitance.

The base and collector current equations of phase II are given by the general charge-control equations, (751.13A) and (751.13B). They are simplified since $q_{BF} >> |q_{BR}|$ because the CB junction is reverse biased by the large collector power supply voltage, V_{CC}. The excess base minority carrier (hole) charge due to this reverse bias, q_{BR}, is very small and negative (deficient or depleted rather than excess charge) - it gives the reverse leakage current of the Shockley diode or the SNS diode. Thus, by setting $q_{BR}=0$ (from $|q_{BR}|<<q_{BF}$), (751.13A) and (751.13B) are simplified and given as follows:

$$j_B(t) \approx dq_{BF}/dt + q_{BF}/\tau_{BF} = J_B u(t) \qquad (753.1A)$$

and

$$j_C(t) \approx q_{BF}/t_{BF} \qquad (753.1B)$$

where u(t) is the unit step function, $u(t<0)=0$ and $u(t\geq 0)=1$. As an exercise of Laplace transform technique, the Laplace transform of the above equations are

$$j_B(s) = sq_{FB}(s) - q_{FB}(t=0^+) + q_{FB}(s)/\tau_{FB} = J_B/s \qquad (753.2A)$$

and

$$j_C(t) = q_{BF}(s)/t_{BF} \qquad (753.2B)$$

where $q_{FB}(t=0+)$ is the initial concentration of the minority carrier charge in the base. Rearranging (753.2A), we get

$$q_{FB}(s) = [(J_B/s) + q_{FB}(0^+)]/(s + \tau_{BF}^{-1}). \qquad (753.3)$$

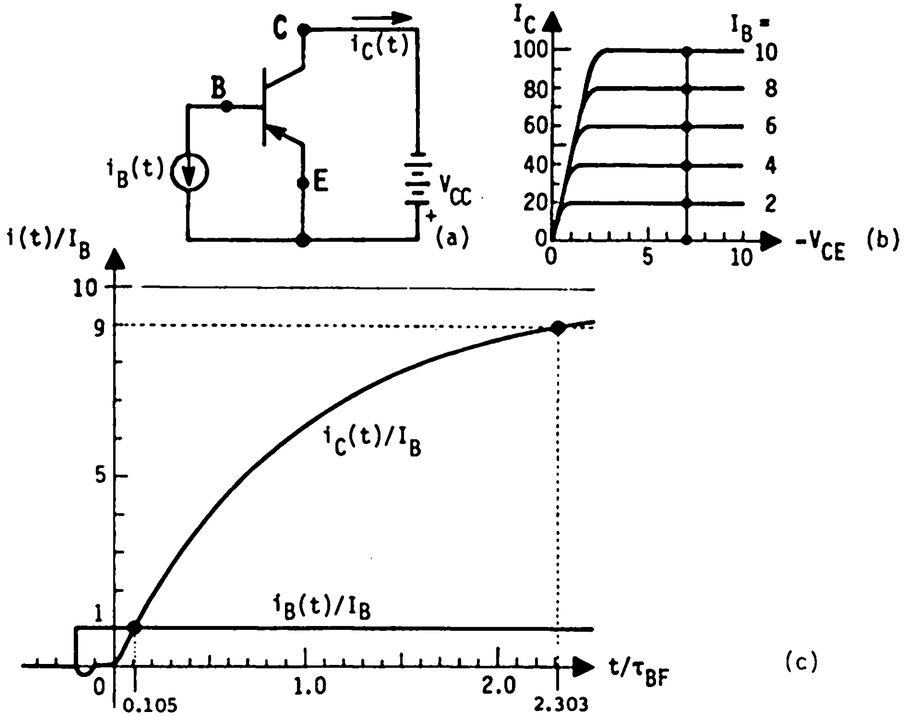


Fig.753.1 Common-emitter BJT switching transient in the active mode due to a base current step drive. (a) The circuit diagram. (b) The quasi-static locus on the output characteristics diagram. (c) The collector current waveform for $\beta_F = 10$.

The solution is obtained by applying the inverse Laplace transform to the above equation. The solution in real time can be looked up from a table of Laplace transform pairs and is

$$q_{FB}(t) = J_B\tau_{BF}[1 - \exp(-t/\tau_{BF})] + q_{BF}(0^+)\exp(-t/\tau_{BF}) \qquad (753.4)$$

$$= J_B\tau_{BF}[1 - \exp(-t/\tau_{BF})]. \qquad (753.4A)$$

(753.4A) is obtained since we assume zero stored excess minority carrier in the base layer, $q_{FB}(t=0^+)=0$. The collector terminal current is then

$$j_C(t) = q_{FB}(t)/t_{BF} = (\tau_{BF}/t_{BF})J_B[1 - \exp(-t/\tau_{BF})]$$

$$= \beta_F J_B[1 - \exp(-t/\tau_{BF})]. \qquad (753.5)$$

During the initial transient for short times, $t << \tau_{BF}$, this reduces to

$$j_C(t) \approx \beta_F J_B t/\tau_{BF} = J_B t/t_{BF}. \qquad (753.5A)$$

Two important results are to be noted: (1) the initial rise is nearly linear and (2) the time required for the collector current to rise to the base drive value, i.e. $j_C(t_1)=J_B$, is $t_1 \approx t_{BF}$, which is obvious from (753.5A). The condition on β_F to give $t_1 \approx t_{BF}$ can also be derived. Using $j_C(t_1)=J_B$ in (753.5), then

$$t_1 = \tau_{BF}\cdot\log_e[1/(1-\beta_F^{-1})]$$

$$\approx \tau_{BF}\cdot[\beta_F^{-1} - \beta_F^{-2}/2 + \ldots] = \tau_{BF}(\tau_{BF}/t_{BF})^{-1} = t_{BF}. \qquad (753.6A)$$

The approximation is valid since β_F is usually large, greater than 10 and frequently exceeds 50 or 100. Thus, this result in words is: **the collector current rises to the base drive value in one diffusion transit time through the base.** This result is also illustrated by the intersection of the $j_B(t)=J_B u(t)$ and $j_C(t)$ waveforms at $t=t_{0.1}=0.105\tau_{BF}$ shown in Fig.753.1(c).

The expanded scale of Fig.753.1(c) is used to show also the initial delay for $j_C(t)$ to reach 10% of its final value, $t_{0.1}$, and the total delay for $j_C(t)$ to reach 90% of its final value. These are

$$t_{0.1} = \tau_{BF}\log_e(1/0.9) = 0.1054\tau_{BF} \qquad (753.6B)$$

and

$$t_{0.9} = \tau_{BF}\log_e(1/0.1) = 2.3026\tau_{BF}. \qquad (753.6C)$$

Although the base charge solution is exact, the collector current is not because of the constant base-transit-time quasi-static approximation which defined the collector current, $j_{CF}(t)=q_{BF}(t)/t_{BF}$, in (751.12B). The two exact solutions can be obtained by solving the diffusion equation using the Laplace transform and separation of variable methods. The Laplace transform solution, with good convergence for short times, is given by

$$j_C(t) \approx J_B\left[\exp[-(X_B/\sqrt{D_B\tau_{BF}})]\cdot \mathrm{erfc}[(X_B/2\sqrt{D_Bt})-\sqrt{t/\tau_{BF}}]\right.$$

$$\left. + \exp[+(X_B/\sqrt{D_B\tau_{BF}})]\cdot \mathrm{erfc}[(X_B/2\sqrt{D_Bt})+\sqrt{t/\tau_{BF}}] + \ldots\right] \qquad (753.7)$$

$$\approx 2J_B\mathrm{erfc}[(X_B/2\sqrt{D_Bt})] \qquad (753.7A)$$

$$\approx (2/\sqrt{\pi})\cdot J_B\cdot\sqrt{(2t/t_{BF})}\cdot\exp(-t_{BF}/2t). \qquad (753.7B)$$

The approximate result at short times is identical to the result of the CB turn-on transient at short times given by (752.35A). Thus, the initial delay, t_1, from the lump model, $0.1054\tau_{BF}$ given by (753.6B), is in serious error compared with the exact solution of $0.26\tau_{BF}$. The lump solution at long times given by (753.6C) is more accurate than the short-time lump solution.

Capacitance Speed-Up of Common-Emitter Turn-On Transient

A constant-voltage base drive in addition to a constant-current base drive can speed up the turn-on transient considerably because the constant-voltage step will give an initial current spike in the collector current which compensates the slowly rising collector current from a base current step just derived. This is demonstrated by the circuit shown in Fig.753.2(a) which gives the collector current waveforms shown in Fig.753.2(b). The base voltage step, $V_1u(t)$, is applied through a series capacitance C_1 to supply a charge drive, $Q_1=C_1V_1$. The base current equation (753.1A) is now

$$j_B(t) \approx dq_{BF}/dt + q_{BF}/\tau_{BF} = J_Bu(t) + C_1d(v_1-v_{BE})/dt \qquad (753.8)$$

$$\approx J_Bu(t) + C_1V_1u_1(t) \qquad (753.8A)$$

where u_1 is the derivative of the unit step function, $u(t)$, and $v_1(t)>>v_{BE}(t)$ is assumed which can be easily attained in practice since v_{BE} is less than one volt. The solution (using Laplace transform) is

$$q_{FB}(t) = \tau_{BF}J_B\{1 - [1 - (C_1V_1/J_B\tau_{BF})]\cdot\exp(-t/\tau_{BF})\} \qquad (753.9A)$$

and

$$j_C(t) = q_{FB}(t)/t_{FB}$$

$$= \beta_FJ_B\{1 - [1 - (C_1V_1/J_B\tau_{BF})]\cdot\exp(-t/\tau_{BF})\} \qquad (753.9B)$$

Thus, if $Q_1=C_1V_1=J_B\tau_{BF}$, then $j_C(t)$ follows truthfully $j_B(t)$ with a gain of β_F without any delay due to diffusion and recombination. In practice, stray capacitance and inductances, current dependent β_F, and base-diffusion transit time will give some delay but significant speed up is still attained by C_1.

Instead of a voltage and a current source just discussed, the collector current response can also be sped up by a base voltage step in series with a RC circuit whose time constant is $R_1C_1 = \tau_{BF}$. The circuit diagram is shown in Fig.753.2(c) and the solution is

$$j_C(t) = \beta_F(V_1/R_1)\{1 - [1 - (R_1C_1/\tau_{BF})]\cdot\exp(-t/\tau_{BF})\} \qquad (753.9C)$$

Note that (753.9C) and (753.9B) are identical if $V_1/R_1 = J_B$.

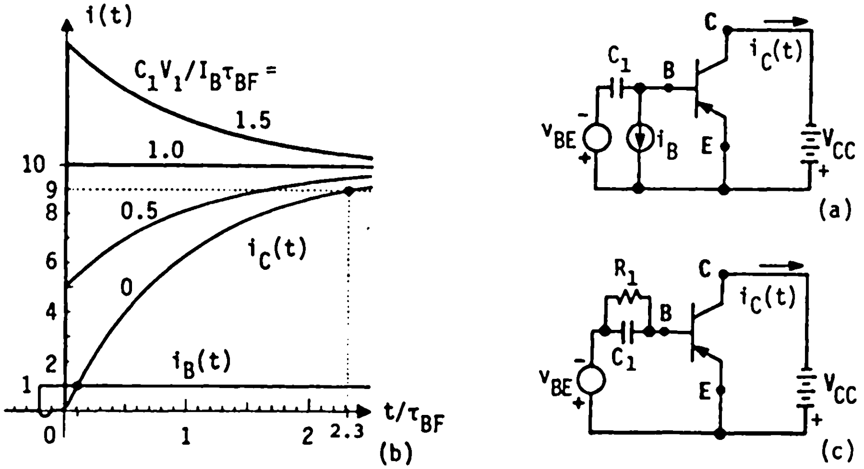


Fig.753.2 Capacitance speed-up of common-emitter BJT switching transient. (a) Circuit to illustrate the principle. (b) Speed-up waveforms. (c) Practical series RC circuit in the base terminal.

Common-Emitter Turn-Off Transient

After the BJT is turned on and reached the steady-state, it is turned off by removing the base drive. Like the CB case, there are two turn-off schemes, open-base to isolate the drive or short-base (short-circuit base-emitter) to divert the drive to ground. The open-base solution is a simple exponential since $j_B(t) = 0$ at all times. Thus,

$$j_C(t) = j_E(t) = \beta_F J_B \exp(-t/\tau_{BF}) \qquad (753.10)$$

The CE short-circuit base-emitter or shorted-base solution is identical to that of the shorted-emitter solution of the CB configuration since the circuits are identical. A substitution of J_E in the CB solutions by $\beta_F J_B$ gives the CE solutions which are β_F times faster, with a time constant given by $t_B = \tau_{BF}/\beta_F$.

754 Comparison of CB and CE BJT Switching Transients

The entire collector current waveform in the CB and CE configurations are summarized in Figs.754.1(a)-(c). Figure (a) shows the four rectangular input waveforms: constant-current open-circuit drives (labeled O) and constant-voltage short-circuit drives (labeled S). The collector current waveforms are given in figure (b) for common-base configuration and in figure (c) for common-emitter configuration. The components of **average propagation delay** (the average of the turn-on and turn-off delays to 50% of the final value, t_{PD}) are labeled in Fig.754.1(d). The contributions from turn-on ($t_{PD\text{-}on}$), turn-off ($t_{PD\text{-}off}$), phase I (t_{PD1}), and phase II (t_{PD2}) are defined by the following general equations:

$$t_{PD} = \tfrac{1}{2}\{[t_{rd1} + (t_{rd2}+\tfrac{1}{2}t_{r2})] + [t_{fd1} + (t_{fd2}+\tfrac{1}{2}t_{f2})]\} \qquad (754.1)$$

$$= \tfrac{1}{2}\{t_{PD\text{-}on} + t_{PD\text{-}off}\} \qquad (754.2)$$

$$= \tfrac{1}{2}\{[t_{rd1}+t_{fd1}] + [(t_{rd2}+\tfrac{1}{2}t_{r2}) + (t_{fd2}+\tfrac{1}{2}t_{f2})]\} \qquad (754.3)$$

$$= \{t_{PD1} + t_{PD2}\}.\ \text{(Note: No ½.)} \qquad (754.4)$$

where

$$t_{PD1} = \tfrac{1}{2}[t_{rd1} + t_{fd1}]$$

and

$$t_{PD2} = \tfrac{1}{2}[(t_{rd2} + \tfrac{1}{2}t_{r2}) + (t_{fd2} + \tfrac{1}{2}t_{f2})]. \qquad (754.5)$$

The $t_{PD\text{-}on}$ and $t_{PD\text{-}off}$ given above and labeled in Fig.754.1(d) are the most commonly given in data sheets. Listed below are the three numerical equations of the gate delay for the two circuit configurations (CB and CE) and the two input drives, constant voltage or constant current (open- or short-circuit sources). The numerical values given in Figs.754.1(b) and (c) are from the exact transient solutions obtained in sections 752 and 753.

Short-Circuit or Voltage Source in CB or CE Configuration (S,CE,CB)

$$t_{PD2} = \tfrac{1}{2}\{[0.26+\tfrac{1}{2}(2.06-0.26)] + [0.13+\tfrac{1}{2}(0.67-0.13)]\}t_B$$

$$= \tfrac{1}{2}\{[1.16] + [0.37]\}t_B = 0.76t_B \qquad (754.6)$$

Open-Circuit or Current Source in CB Configuration (O,CB):

$$t_{PD2} = \tfrac{1}{2}\{[1.16] + [1.16]\}t_B = 1.16t_B \qquad (754.7)$$

Open-Circuit or Current source in CE Configuration (O,CE)

$$t_{PD2} = \tfrac{1}{2}\{[0.105+\tfrac{1}{2}(2.303-0.105)] + [0.105+\tfrac{1}{2}(2.303-0.105]\}\tau_B$$

$$= \tfrac{1}{2}\{[1.204] + [1.204]\}\tau_B = 1.20\tau_B \qquad (754.8)$$

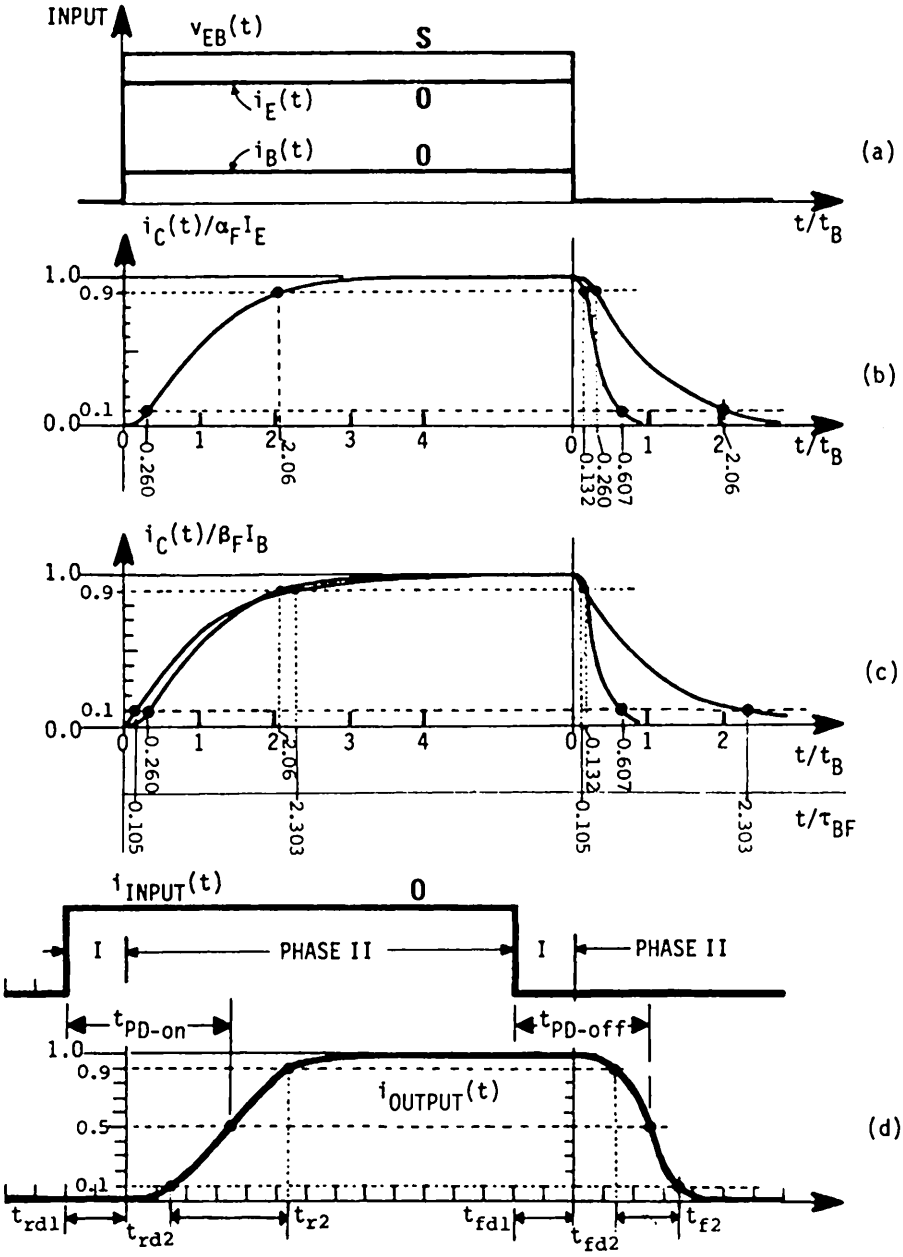


Fig.754.1 Comparison of BJT transient responses in CB and CE configurations. (a) Input waveforms. Output (collector current) waveforms: (b) CB with t/t_B scale, and (c) CE with t/t_B and t/τ_B scales. (d) Definition of average propagation delay. O=Open-circuit input drive. S=Short-circuit input drive. As an exercise, label (d) according to (754.6)-(754.8).

The phase I delays were shown to come from charging the emitter-base and collector-base junction space-charge layer capacitances by the majority carriers. The phase II delays were shown to originate from minority carrier diffusion through the base layer and recombination in the base layer. The two phases are sharply separated in early generation BJTs which were low-frequency and slow devices due to a slow phase II from thick base and large lifetime. Such separation enables succinct and clear pedagogical description of the physics. In today's (1990) state-of-the-art submicron BJTs, the phase II minority-carrier diffusion delay is shortened towards 1ps. This is comparable to or less than the phase I delays from charging the space-charge-layer capacitances by majority carriers. Thus, the two phases are no longer distinct and analytical solutions valid for all times are no longer possible. However, the foregoing analytical results can still be used to provide design guide. Numerical examples will be given in the next section, 755, illustrating results of new technologies to speed up the BJTs.

Equations (754.6) to (754.8) show that the main difference between the two circuit configurations and two drives is the much longer time constant of the CE open-circuit or current drive (O,CE), (754.8) which is controlled by minority-carrier recombination lifetime in the base, $\tau_B(\geq 1\text{ns}=1000\text{ps})$. The characteristic time of the other three cases, (O,CB), (S,CB) and (S,CE) given by (754.6) and (754.7), is controlled by the much shorter minority-carrier diffusion delay in the base, $t_B(\approx 1\text{-}100\text{ps})$. The key physics underlying the difference is the necessity of storing (putting in) and removing charges via recombination when the base lead is open in the (CE,O) circuit, $i_B=0$, because this open-circuits both the collector-base and emitter-base junctions. The open junctions are not able to inject minority carriers into and sweep them out of the base layer, thus, the temporal change of the minority carrier concentration can only be effected via the slow recombination. In contrast, when one or both junctions are short-circuited as in the three cases (O,CB), (S,CB), and (S,CE), the carrier concentration change is effected by the very fast carrier diffusion through the base to or from the short-circuited junctions.

When base diffusion delay is reduced to less than 10ps, delays due to two sources must be included which increase phases I and II and the total delays. These are: (i) charging the space-charge layer capacitances through the base resistance, and (ii) the drift transit time through the collector-base space-charge layer. They are discussed in the following section.

755 Speeding Up the BJT via Technology

The BJT switching transients can be shortened by two means: improving the intrinsic speed of the transistor to be discussed in this section, and reducing the circuit delays to be discussed in the following sections, 76n.

In the preceding sections, it is shown that the intrinsic switching speed of a BJT, as measured by the delay of the collector current or collector voltage across a small resistance in response to a large-signal input current or voltage transient applied to the base (CE) or emitter (CB), are limited by the following sequence of events in a p/n/p BJT. The load resistance is assumed zero so that it does not introduce additional delay. Note also that the small-signal transient response is different and faster as expected and indicated in the next paragraph.

(1) Charging the emitter-base (CB) or base-emitter (CE) junction space-charge layer capacitance by the majority carrier to a sufficiently large forward bias in order to begin injection of minority carriers into the quasi-neutral emitter and base layers from the other side, (752.12),

$$t_{1\text{O}} = t_{EBJ} \approx 0.378 t_{EB-FB}$$

$$= 6.73(\text{nA-}\mu\text{s})/I_E(\text{nA}) \qquad \text{(Current Source)} \qquad (755.1A)$$

and

$$t_{1S} = r_{b'}(C_{eb0}+C_{eb5}) \qquad \text{(Voltage Source).} \qquad (755.1B)$$

(2) Charging the collector-base junction capacitance by the majority carriers which occurs simultaneously with (1), either via the base spreading resistance in the CB configuration given by (752.22),

$$t_{2CB} = t_{CBJ-CB} = r_{b'}C_{cb5} \qquad \text{(CB Current Source)} \qquad (755.2A1)$$

$$= r_{b'}(C_{cb5}+C_{eb0}) \qquad \text{(CB Voltage Source)} \qquad (755.2A2)$$

or directly via the base and collector spreading resistances in CE configuration which occurs also simultaneously with (1),

$$t_{2CE} = t_{CBJ-CE}$$

$$= r_{c'}[C_{cb5}C_{eb0}/(C_{cb5}+C_{eb0})] \qquad \text{(CE Current Source)} \qquad (755.2B1)$$

and

$$\approx r_{b'}(C_{eb0}+C_{cb5}) + r_{c'}C_{cb5} \qquad \text{(CE Voltage Source).} \qquad (755.2B2)$$

(3) **Minority carrier** (electrons in p/n/p) diffusion time through the quasi-neutral emitter (negligible when this component of current is small, i.e., β_F is high),

$$t_3 = (1-\gamma_F)t_E = (1-\gamma_F)X_E^2/2D_E. \qquad (755.3)$$

(4) **Minority carrier** (holes in p/n/p) diffusion time through the base, which gives a large-signal switching delay of

$$t_4 = \gamma_F X_B^2/2D_B \approx X_B^2/2D_B \equiv t_B. \qquad (755.4)$$

(5) **Minority carrier** (holes in p/n/p) recombination lifetime in the base which gives a large-signal switching delay of

$$t_5 = \tau_B \approx \beta_F t_B. \qquad (755.5)$$

(6) Minority carrier (holes in p/n/p) drift delay through the collector-base junction space-charge layer,

$$t_{CBsat} = X_{CB}/\theta_{sat}. \qquad (755.6)$$

The drift velocity of the minority carriers (electrons) injected into the space-charge layer, θ_{sat}, saturates to about 10^7cm/s at about 10→30kV/cm as indicated by Fig.314.1. (Note the factor of 1/2 in $t_{CBsat}=X_{CB}/2\theta_{sat}$ is present in small-signal responses such as computing f_α and f_t.)

The total **pulse delay**, t_{pd}, to be defined as the **sum of turn-on and turn-off delays to reach 90% of their final values**, is given by the sum of all the terms above, taking into account that the majority and minority carrier delays may overlap. The three fastest circuits among the four are: (i) the CE and CB driven by voltage source with $0.76t_B$ average propagation delay given by (754.6) and $2.66t_B$ total pulse delay (90% turn-on plus turn-off = 2.06+0.607 = 2.667), and (ii) the CB driven by a current source with $1.16t_B$ average propagation delay given by (754.7), and $4.12t_B$ total pulse delay (90% turn-on plus turn-off = 2.06+2.06 = 4.12). The slowest circuit is the CE driven by a current source with average propagation delay of $1.204\tau_B = 1.204\beta_F t_B$ and pulse delay of $(2.303+2.303)\tau_B = 4.606\beta_F t_B$ so that it is $\beta_F(\approx 100)$ times slower than the other three circuits. (In practical CE current source circuits with cascaded BJTs, the driving BJT has a small source resistance so it acts like a voltage source but the driven BJT has a large external input base resistance, so the current source is really a voltage source in series with a base resistance). The **pulse delay**, t_{pd}, for all four circuits is then

$$\begin{aligned} t_{pd} &= (\text{larger of } t_1,\ t_2) + t_3 + K_4t_4 + K_5t_5 + t_6 \\ &= (\text{larger of } t_1,\ t_2) + (1-\gamma_E)t_E + K_d t_B + K_\tau \tau_B + X_{CB}/\theta_{sat} \qquad (755.7) \end{aligned}$$

where the appropriate K_d is to be used for the particular circuit selected and delay definition desired: K_d(50%,90%) = (0.78,2.73) for (S,CB) and (S,CE); and (1.16,4.12) for (O,CB). For (0,CE) K_d can be set to zero since τ_B dominates, and K_τ(50% or 90%) = (1.204,4.606). K_d can be reduced significantly by a built-in electric field from grading the base dopant impurity concentration profile. For concentration drop of 7.4x (7 times) from emitter to collector to speed-up the turn-on transient, the speed-up factor of the base diffusion delay is $\eta=\log_e 7.4=2$. However, this gives an opposing field that delays the sweep-out of a part of the stored base charge which is extracted by the short-circuited emitter-base junction during turn-off in the short-circuit or voltage-source configuration. Thus, taking this into account and for $\eta=2$, the improved K_d (50%,90%) are: $(0.476,1.708)t_B$ for (S,CB) and (S,CE); and $(0.58,2.06)t_B$ for (O,CB), about a 2x speed-up.

In comparison, the small-signal **maximum frequency of oscillation**, f_{Max}, also depends on the six delay mechanisms. However, it has a fundamental difference sometimes overlooked: it is a small-signal property of the BJT which is already biased to a d.c. steady-state point. Thus, the capacitances and the 3-db cutoff frequencies from small-signal delays of minority carrier diffusion in the emitter and base layers are well-defined constants. From (742.9), f_{Max} is

$$f_{Max} = \sqrt{f_{t-}/8\pi r_{b'}C_{c4}} \le \sqrt{f_{\alpha}/8\pi r_{b'}C_{c4}} \qquad (755.8)$$

where C_{c4} is the collector-base junction space-charge layer capacitance, C_{cbt}, at a reverse bias voltage of 3.2V for a built-in voltage of 0.8V and f_t from (742.8C) is

$$\begin{aligned} 1/2\pi f_{t-} &= t_1 + t_2 + t_3 + t_4 + t_6/2 + r_{b'}C_c \\ &= 0 + 0 + (1-\gamma_{eo})t_E + t_B/1.22 + X_{CB}/2\theta_{sat} + r_{b'}C_c \qquad (755.9) \end{aligned}$$

where t_1 and t_2 are already included in t_3, and $t_3 = (C_e+C_{ebt})/g_b$ in (742.8C) is approximated by $(1-\gamma_{eo})t_E$. It is small and can be neglected at high d.c. current since $g_b=qI_C/kT$ is large.

Equation (755.7) shows how to speed up the BJT and (755.9) shows how to increase the f_{Max} of the BJT by varying its geometry and material parameters. The slowest circuit (CE driven by current source) was used in the first mass-produced logic circuits since 1965. Since mid-1985, the three fast circuits have been used in logic products and as the benchmark test of the latest submicron technology advances. The technology to control the material and geometry factors in order to speed up the BJT are discussed in the following two subsections.

Technology of Reducing Recombination Lifetime τ_B

Historically, the slowest circuit [CE driven by a current source given by (754.8)], known as the **transistor-transistor logic** (TTL), was used in the first large volume production of digital logic gates. TTL has dominated digital applications for twenty years (1965-1985) since Thomas A. Longo invented it at Sylvania Electric Co. in 1962 and Texas Instruments cased the design and began volume production during 1965-1966. This was because its simple geometry lending to high yield, low cost and high profit margin. The TTL circuit will be described in section 765.

As indicated by the K_5t_5 $(=K_\tau \cdot \tau_B = 1 \cdot \beta_F t_B = \beta_F t_4)$ in (755.7), the switching delay of the CE current-source circuit used in TTL is mainly determined by recombination in the base layer to remove the minority carriers stored in the quasi-neutral base layer. Thus, the technology developed to speed up the TTL circuit was to reduce the base recombination lifetime. The recombination lifetime in the quasi-neutral base, $t_5=\tau_B$ in (755.5), is given by (372.3) for electrons in a n/p/n BJT and (372.4) for holes in a p/n/p transistor,

$$\tau_B = \tau_n = 1/(c_n^t N_{TT}) \qquad \{\text{n/n/p BJT}\} \qquad (755.10A)$$

$$\tau_B = \tau_p = 1/(c_p^t N_{TT}) \qquad \{\text{p/n/p BJT}\} \qquad (755.10B)$$

$c_n{}^t$ and $c_p{}^t$ are the thermal capture rates of electrons and holes at the recombination center. N_{TT} is the concentration of the recombination in the base layer. The two parameters are discussed separately in the following paragraphs.

Lifetime reduction in Si BJTs has been successfully implemented in the production line of bipolar logic integrated circuits, by incorporating the gold impurity into the Si wafer via solid state diffusion in the final high-temperature fabrication step. Prior experimental research has shown that gold is a very efficient recombination center in silicon. It was also a very reproducible recombination center because of two properties: high diffusivity to give areal and volume uniformity, and sufficiently high solid solubility in Si to give low lifetime. The diffusivity of gold in silicon is very large, $\geq 100\mu m/\sqrt{}hr$ in 800-1200C as indicated in Fig.512.3. Thus, its concentration reaches the maximum value, the solid solubility of gold in Si at the diffusion temperature, in a short diffusion time. Its concentration is practically constant over the entire Si wafer after only a short diffusion time at not-too-high a temperature (in the 800-1000C range). This provides uniformity and equal gold concentration and hence equal lifetime among many 3-4-inch Si wafers in a production batch. The experimentally measured solubility values from the 1960 compilation of F.A.Trumbore of the Bell Telephone Laboratories are:

$$\begin{aligned}&(\text{Concentration cm}^{-3},\ \text{Diffusion temperature C=Celsius}) =\\ &(3.0\times10^{14},\ 700),\ (1.0\times10^{15},\ 800),\ (1.7\times10^{15},\ 850),\ (3.1\times10^{15},\ 900),\\ &(5.5\times10^{15},\ 950),\ (1.0\times10^{16},1000),\ (1.7\times10^{16},1050),\ (2.8\times10^{16},1100),\\ &(7.0\times10^{16},1200),\ (1.2\times10^{17},1200).\end{aligned} \tag{755.11}$$

To estimate the lowest practical recombination lifetime via gold center, the thermal capture rate coefficient of electrons and holes at the gold center, $c_n{}^t$ and $c_p{}^t$, must be known. Gold has two energy levels in Si and only one is the dominant active recombination level for minority carriers. In p-type Si or quasi-neutral p-base of an Si n/p/n transistor, most of the gold centers are occupied by holes due to the abundance of holes in p-Si, hence, the gold centers are positively charged and in the donor charge state. Thus, the active recombination energy level is the lower-gap donor level, E_V+345meV shown in Fig.381.2, which is empty (or occupied by a hole) and ready to capture an injected electron (minority carrier). In n-type Si or n-base of the p/n/p transistor, the dominant hole capture level is the midgap acceptor level, E_C-547meV shown in Fig.381.2. It is occupied by an electron and negatively charged because of the abundance of electrons in n-type Si base layer.

The thermal capture rates of minority carriers at the gold recombination center, $c_n{}^t$ and $c_p{}^t$, can be computed from the emission rate data (Fig.381.2) using the mass action law, $e_n{}^t = c_n{}^t n_1 = c_{nt}N_C\exp[(E_C\text{-}E_T)/kT]$ or $c_n{}^t = (e_n{}^t/N_C)\exp[-(E_C\text{-}E_T)/kT]$. However, the computed value may be in error by a factor of ten or more because the emission rates are measured at high electric fields instead of equilibrium (The electric field in the quasi-neutral base is zero or very low.) and because the measured thermal activation energy from the slope of the emission-rate versus 1/T data (Fig.381.2) may differ from the true energy level, E_C-E_T, by several (kT/q)'s. Thus, thermal capture rate data, accurate enough for BJT design purposes, must be measured directly rather than computed from the thermal emission rate data. Accurate capture rates have indeed been measured using the junction capacitance and current transient technique pioneered by this author and his doctoral graduate students during 1964-1987 (Tasch, Rosiser, Forbes, Yau, Herman, Walker, Wang, Jackson, and Lu). Nevertheless, we shall give a simple and yet accurate back-of-the-envelope estimate using the simple Bohr hydrogen model described in (381.5) in order to bring out the fundamental physics underlying the quantum mechanical transition of electron and hole capture by the gold potential well. This illustration will also give a method to estimate the capture rates at other impurity and defect centers, few of which have been measured accurately. The Bohr hydrogen energy level and orbit formulae given in section 223 are

$$E_n = -m^*q^4/[2\hbar^2(4\pi\varepsilon_s)^2n^2] = -13.6050(m^*/m)(\varepsilon_o/\varepsilon_s)^2n^{-2} \text{ eV}$$

and

$$a_n = (4\pi\varepsilon_s)\hbar^2n^2/m^*q^2 = 0.5292(m/m^*)(\varepsilon_s/\varepsilon_o)n^2 \text{ A.} \qquad (755.12)$$

Eliminating $(m/m^*)n^2$ between the two expressions, then

$$E_na_n = q^2/2(4\pi\varepsilon_s) = 7.20(\text{eV-A})(\varepsilon_o/\varepsilon_s). \qquad (755.13)$$

Both m^* and ε_s can be adjusted to fit the data, but m^* is closer to m (just like the electron in hydrogen) than ε_s to ε_o which varies from 11.7 to 1 when the orbit shrinks and fewer valence electrons are enclosed to screen the Coulomb force of the ionized gold atom. Thus, ε_s is the adjustable parameter. Let $n=1$, $m=m^*$, and use a thermal velocity of 2×10^7cm/s for both electrons and holes, then

$$\varepsilon_s/\varepsilon_o = \sqrt{(13.605/E_1)} \qquad (755.14A)$$

$$a_1 = 0.5292\sqrt{(13.605/E_1)} = \sqrt{(3.810/E_1\text{eV})} \text{ A} \qquad (755.14B)$$

$$\sigma^t = \pi a_1{}^2 = 3.810\pi/E_1 \text{ A}^2 = [1.197/E_1(\text{eV})]\text{x}10^{-15} \text{ cm}^2 \qquad (755.14C)$$

and

$$c^t = \sigma^t\theta_{th} = \sigma^t\cdot 2\text{x}10^7 = [2.394/E_1(\text{eV})]\text{x}10^{-8} \text{ cm}^2 \qquad (755.14D)$$

Thus, for electron-capture into the ground state of the positively charged donor level at E_V+345meV = E_C-[(590+547)-345] = E_C-792meV, we have:

$$\varepsilon_s/\varepsilon_o = \sqrt{(13.605/E_1)} = \sqrt{(13.605/0.792)} = 4.144 \qquad (755.15A)$$

$$a_1 = \sqrt{(3.810/E_1)} = \sqrt{(3.810/0.792)} = 2.194\text{A} \qquad (755.15B)$$

$$\sigma^t_{n+1} = \pi a_1{}^2 = \pi\cdot 2.194^2 = 15.12 \text{ A}^2 = 1.512\text{x}10^{-15}\text{cm}^2 \qquad (755.15C)$$

$$c^t_{n+1} = \sigma^t_{n+1}\theta_{th} = 1.512\text{x}10^{-15}\text{x}2\text{x}10^7 = 3.0\text{x}10^{-8}\text{cm}^3/\text{s} \qquad (755.15D)$$

and for hole-capture into the negatively charged gold donor center at E_V+590meV, we have

$$\varepsilon_s/\varepsilon_o = \sqrt{13.605/0.590} = 4.802 \qquad (755.16A)$$

$$a_1 = \sqrt{3.810/0.590} = 2.541A \qquad (755.16B)$$

$$\sigma^t_{p-1} = \pi\cdot 2.541^2 = 20.28A^2 = 2.028\mathrm{x}10^{-15}\mathrm{cm}^2 \qquad (755.16C)$$

and

$$c^t_{p-1} = 2.028\mathrm{x}10^{-15}\mathrm{x}2\mathrm{x}10^7 = 4.0\mathrm{x}10^{-8}\mathrm{cm}^3/\mathrm{s}. \qquad (755.16D)$$

The above values of the capture cross section and rate are within a factor of two to three of the experimental measurements. It illustrates the power of using simple physics to get accurate results.

Using the solid solubility at a gold diffusion temperature of 1000°C ($N_{TT}=N_{Au}=1.0\mathrm{x}10^{16}\mathrm{cm}^{-3}$) and capture rates given above, the recombination lifetimes in the base layer are

$$\tau_B = \tau_n = 1/(c^t_{n+1}N_{TT}) = 1/(3.0\mathrm{x}10^{-8}\mathrm{x}10^{16}) = 3.3\mathrm{ns} \quad \{\mathrm{n/p/n\ BJT}\} \qquad (755.17A)$$

$$\tau_B = \tau_p = 1/(c^t_{p-1}N_{TT}) = 1/(4.0\mathrm{x}10^{-8}\mathrm{x}10^{16}) = 2.4\mathrm{ns}. \quad \{\mathrm{p/n/p\ BJT}\} \qquad (755.17B)$$

Gold was used initially for many years to increase the TTL speed but due to its many limitations, device design (Schottky-barrier Bethe diode clamp) and circuit innovations (ECL) have been developed to speed up the TTL gate. The limitations are as follows. (1) Au cannot be diffused higher than about 1000C to give a higher Au concentration because B, As, and P diffusion at the higher temperatures would move the base and emitter junctions. Thus, τ_B cannot be reduced to less than about 1-ns in Si by increasing gold concentration. Other impurities (such as Ag, used in p/n/p/n 4-layer power control rectifier) may give a lower τ_n but much higher τ_p so that they cannot simultaneously reduce τ_B in n/p/n and p/n/p on the same chip. (2) Except Au, it is difficult to introduce a high concentration of impurities into Si uniformly without generating defects. Defects would increase the leakage current and lower the junction breakdown voltage. (3) Au reduces β_F in both n/p/n and p/n/p, but more severely in p/n/p that it was difficult to give reproducible results by thinning the base to compensate for the drop in β_F.

For the midgap Au acceptor level at 300K shown in Fig.381.2, $e_{p0}{}^t = 1/(1\mathrm{x}10^{-2}) = 100/\mathrm{s}$ and $e_{n1}{}^t = 1/(1\mathrm{x}10^{-3}) = 1000/\mathrm{s}$. Let $N_{TT} = 10^{16}\mathrm{Au/cm}^3$ and a depleted space-charge-layer volume of $\mathrm{Area}\cdot X_{pn} = 1\mu\mathrm{m}^2\cdot 1\mu\mathrm{m} = 1\mu\mathrm{m}^3$, then the leakage current due to e-h generation from the Au center in Si is

$$\begin{aligned} I &= q\cdot[\text{generation rate}]\cdot[\text{number of generation centers}] \\ &= q\cdot[e^t_n e^t_p/(e^t_n+e^t_p)]\cdot[N_{TT}\cdot\mathrm{Area}\cdot X_{pn}] \\ &= 1.6\mathrm{x}10^{-19}\cdot[1000\mathrm{x}100/(1000+100)]\cdot[10^{16}\mathrm{x}10^{-8}\mathrm{x}10^{-4}]\mathrm{A} = 145\mathrm{fA}. \end{aligned}$$

For a VLSI chip with $2\mathrm{mm}\cdot 5\mathrm{mm} = 1\mathrm{x}10^7\ \mu\mathrm{m}^2$ junction area and a $1\mu\mathrm{m}$ thick space-charge layer, the leakage current is $145\mathrm{fAx}10^7 = 1.45\mu\mathrm{A}$. This shows that the much higher observed leakage current is due to Au-diffusion induced defects rather than Au as a generation center alone.

Speeding-Up via Geometry and Resistivity Reduction

When the BJT is driven by a short-circuit or voltage source in the CB or CE configuration, (S,CB) and (S,CE), or by an open-circuit or current source in the CB configuration, (O,CB), minority carrier recombination delay in the base ($\tau_B = 1$ns) is no longer an important factor because the stored minority carriers are swept out of the base by the short-circuited collector-base and emitter-base junctions in one to two base-diffusion transit time, $t_B = X_B^2/2D_B$. This is in the range of 1-10ps which is 100 to 1000 times smaller than the minority carrier recombination lifetime, $\tau_B \gtrsim 1$ns. Note $\tau_B \gtrsim \beta_F t_B$ or $\beta_F \lesssim \tau_B/t_B$ at high J_C from (738.33B). Thus, the base-layer thickness may be thinned to reduce t_B and increase speed. However, the base spreading resistance would also increase due to thinner base, which would increase $r_{b'}C_{cb}$ and lower the maximum frequency of oscillation.

A novel technology was developed during the 1980's, known as the **self-aligned polysilicon-emitter bipolar technology**. It gives emitter widths smaller than the lithographic linewidth and reduces the collector-base underlap diode area. A second technology, **dielectric isolation**, was used to isolate the transistor, further reducing the collector-base underlap diode area. Figure 755.1 shows the top and cross-sectional views of a conventional non-self-aligned BJT on the left (See also Fig.741.2.) and a 1990 **self-aligned** BJT on the right. A comparison of these two structures is described in the following paragraphs.

Figure 755.1 shows that the spacing between the emitter and base contacts, W_{EB}, cannot be made less than about twice the lithographic linewidth in the non-self-aligned structure. This gives a large base resistance, $r_{b'} = \rho_B L_{Tb'}/(2X_B L_E)$, because of the long length of the base resistance, $L_{Tb'}$. The self-aligned structure shown in the figure closes this emitter-base contact gap, $W_{EB}=0$. Consequently, $L_{Tb'}$ is reduced from twice to less than half of the lithographic linewidth and $r_{b'}$ is reduced by more than 4 times. Lateral base diffusion of the p+ boron source towards the n+ poly-Si emitter contact can further reduce the emitter-base junction width substantially below the lithographic linewidth. For example, $W_E = 0.35\mu$m using 1-μm lithography giving $L_{Tb'} = W_E/2 = 0.175\mu$m. The heavily As-doped n-type polycrystalline silicon emitter is used for three purposes: (1) self-aligned n-emitter contact, (2) n-emitter impurity source to give high β_F consistently without generating defects and As or P impurity-clusters that give CE pipes which short-circuit the base layer, and (3) self-aligned implant and diffusion mask for defining the self-aligned boron-doped poly-Si contact to the p+base. The boron-doped polysilicon base contact is used as an additional impurity source to increase the boron concentration in the base contact area. In practice, the **double poly** Si layers overlap and are separated by an oxide thermally grown on the lower poly-Si layer.

The area of the collector-base underlap diode and hence C_{cb} are reduced by self-alignment described above. They are further reduced by a second technology,

dielectric isolation, which uses regrown and deposited oxide in a ring trench etched into the Si surface, known as **recessed oxide**. It is indicated by the two vertical trenches of the self-aligned structure in Fig.755.1. These two geometry innovations have reduced the delay due to $r_b{}'C_{cb}$ by a factor of nearly ten and sped up BJTs to approach the intrinsic speed around 1ps.

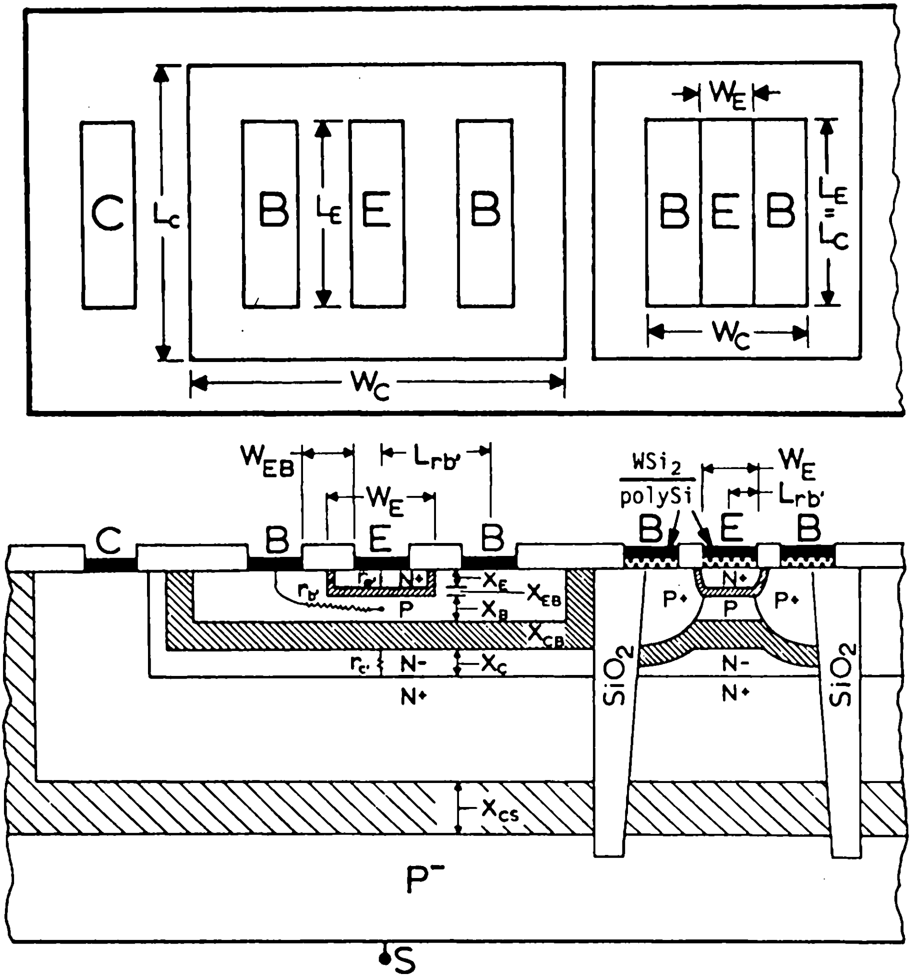

C
L_C
B
L_E
E
B
W_C
W_E
B
E
B
$L_E = L_C$
W_C
W_{EB}
$L_{rb'}$
W_E
C
B
E
B
$r_{b'}$
$r_{e'}$
N+
X_E
X_{EB}
P
X_B
X_{CB}
$r_{c'}$
N-
X_C
N+
WSi_2
polySi
W_E
$L_{rb'}$
B
E
B
N+
P+
P
P+
SiO_2
N-
N+
SiO_2
X_{CS}
P^-
S

Fig.755.1 The top and cross-sectional views of two BJTs; one without (left) and one with (right) the self-aligned emitter and dielectric isolation. See also Fig.741.2 for the left figure.

Table 755.1
Performance Comparison of Four Micron-Submicron Si BJTs

TRANSISTOR NUMBER			(1)	(2)	(3)	(4)
TECHNOLOGY			STANDARD	STANDARD	SELF-ALIGN	SELF-ALIGN
LINEWIDTH	(μm)		1.0	1.0	1.0	0.5
W_{EB} E-B GAP	(μm)	n+	1.0	1.0	0.0	0.0
$N_{EE}(\Delta E_G=60mV)$	(cm^{-3})	n+	$1.0x10^{20}$	$1.0x10^{20}$	$<1.0x10^{20}$	$<1.0x10^{20}$
D_E(holes)	(cm^2/s)		2	2	2	2
W_E(width)	(μm)	n+	1.0	1.0	1.0	0.5
$L_E=L_B$(length)	(μm)	n+	5	5	5	5
$X_{Ej}=X_E$	(μm)		1.0	0.1	0.1	0.1
P_{BB} (0.1Ω-cm)	(cm^{-3})	p	$5.0x10^{17}$	$5.0x10^{17}$	$5.0x10^{17}$	$5.0x10^{17}$
D_B(electrons)	(cm_2/s)		10	10	10	10
V_{EB-bi}(nodeg)	(V)	p	1.053	1.053	1.053	1.053
X_{EB-bi}	(μm)	p	0.051	0.051	0.051	0.051
$C_{eb1}(@V_B=1V)$	(fF)	p	10.2	10.2	10.2	5.1
X_B	(μm)	p	1.0	0.10	0.10	0.10
$L_{rb'}$(length)	(μm)	p	2.0	2.0	0.5	0.25
R_{Bsq}(0.1Ω-cm)	(Ω/sq)	p	1k	10k	10k	10k
$r_{b'}$	(Ω)	p	200	2000	500	250
X_{Bj}	(μm)		2.051	0.251	0.251	0.251
N_{CC}(0.15Ω-cm)	(cm^{-3})	n-	$5.0x10^{16}$	$5.0x10^{16}$	$5.0x10^{16}$	$5.0x10^{16}$
W_C(width)	(μm)	n-	7.0	7.0	5.0	3.0
L_C(length)	(μm)	n-	7.0	7.0	5.0	5.0
V_{CB-bi}	(V)	n-	0.857	0.857	0.857	0.857
X_{CB-bi}	(μm)	n-	0.1686	0.1686	0.1686	0.1686
$C_{cb1}(@V_B=1V)$	(fF)	n-	30.1	30.1	15.4	9.24
X_C	(μm)	n-	0.782	0.33	0.33	0.33
$r_{c'}$	(Ω)	n-	230	99	99	198
X_{cj}(epit n^-)	(μm)		3.051	0.751	0.751	0.751
X_{N+}(buried)	(μm)	n+	1	1	1	1
P_{SUB}(3Ω-cm)	(cm^{-3})		$5.0x10^{15}$	$5.0x10^{15}$	$5.0x10^{15}$	$5.0x10^{15}$
W_{SUB}	(μm)	p-	11	11	7	3.5
L_{SUB}	(μm)	p-	9	9	5	5
X_{CS1}(n+/p-)	(μm)	p-	0.553	0.553	0.553	0.553
C_{cs1}	(fF)	p-	19.2	19.2	6.8	3.4
$t_E(=X_E^2/2D_E)$	(ps)		2500	25	25	25
$r_{b'}C_{eb1}$	(ps)		2.04	20.2	5.0	1.25
$\tau_B(=10ns)$	(ps)		10k	10k	10k	10k
$t_B(=X_B^2/2D_B)$	(ps)		500.0	5.0	5.0	5.0
$r_{b'}C_{cb1}$	(ps)		6.02	60.2	7.70	2.31
X_{CB4}/θ_{sat}	(ps)		3.37	3.37	3.37	3.37
$1-\gamma_{eo}$			0.01	0.01	0.01	0.01
$1-\alpha_{bo}$			0.05	0.0005	0.0005	0.0005
$\beta_F=\beta_{fo}$			16.6	95.2	95.2	95.2
$(1-\gamma_{eo})t_E$	(ps)		25.0	0.25	0.25	0.25
t_{pd}(90% pulse-delay)	(ps)		1395.	34.5	20.5	18.1
$f_T=f_{bw}(\beta=h_{21}=1)$	(GHz)		0.358	26.4	26.4	26.4
$f_{T-}(\text{for } f_{Max}) < f_T$	(GHz)		0.356	4.42	16.2	22.1
$f_{Max}=\sqrt{f_{T-}/8\pi r_{b'}C_{cb4}}$	(GHz)		2.2	2.41	12.9	27.6

Table 755.1 lists the numerical results computed for four BJTs to delineate the factors that give the improvement of speed. Transistors (1) and (2) are not self-aligned while (3) and (4) are. Transistors (1), (2) and (3) use the 1-micron lithography while transistor (4), the 0.5-micron lithography. Finally, the emitter and base layer thicknesses are 1-μm in (1) and 0.1μm in (2), (3) and (4). Formulae used are from (755.1) to (755.9) and listed later.

Comparing transistors (1) and (2), one notes that the base diffusion delay, t_B, is decreased 100 times from 500ps to 5ps by making the base 10 times thinner, from 1-μm to 0.1-μm, since $t_b \propto X_B^2$. However, the base resistance is increased 10 times from 200Ω to 2000Ω because $r_{b'} \propto X_B^{-1}$. Thus, the $r_{b'}C_{cb}$ time constant is increased 10 times from 6.02ps to 60.2ps. (The significant figure given is to help readers to locate the number in the table.)

Transistors (3) and (4) have the same emitter area, layer thicknesses and dopant impurity concentration profiles. The only difference is that (3) employs the self-alignment emitter and dielectric isolation technologies. Self-alignment reduces the base resistance 4x (4 times) from 2000Ω to 500Ω because the effective length, $L_{rb'}$, is now 0.5μm instead of twice the lithographic linewidth, 2μm. Combined self-alignment and dielectric isolation reduce the collector area 50% from 7x7μm^2 to 5x5μm^2. Then $r_b'C_{cb}$ delay is reduced 8x from 60.2ps to 7.70ps.

Transistor (4) is a downsized version of (3) from 1-μm to 0.5-μm lithography. $r_{b'}$ is reduced 2x from 500Ω to 250Ω. The collector-base junction area is also reduced, by 40%, from 5x5μm^2 to 5x3μm^2 and the C_{cb} is reduced by the same amount. Thus, $r_b'C_{cbl}$ is reduced by $1/0.3=3.33$ from 7.70ps to 2.31ps.

When the base transit time, t_B, and collector capacitance charging time, $r_{b'}C_{cb}$, are reduced to less than 10ps, the transit delay time through the collector-base junction space-charge layer at the phonon-scatttering limited velocity, X_{CB}/Θ_{sat}, becomes significant. The above transistors have $X_{CB4}/\Theta_{sat}=3.37$ps at a barrier height of 4V or about 3.2V reverse collector-base junction voltage which gives a $X_{CB4}=2X_{CB\text{-}bi}=0.3372\mu$m-thick depletion layer. The 3.37ps X_{CB}-transit-time delay is very significant in transistors (3) and (4) where $t_B=5$ps and $r_{b'}C_{bc} \approx 7.7$ps and 2.3ps.

The four figures-of-merit are given in the last four lines of Table 755.1. These are: t_{pd} (pulse delay time), $f_t \approx f_{bw}$ given by (744.7), f_{t-} given by (742.8C) for computing f_{Max}, and f_{Max} given by (742.9). It is evident that transistor (1) is limited by the diffusion delay through the 1-micron thick base, $t_B=500$ps, resulting in $t_{pd}=1395$ps, $f_{bw}(\approx f_t)=0.358$GHz, $f_{t-}=0.356$GHz and $f_{Max}=2.2$GHz. When the base is thinned to 0.1-micron in transistor (2) without self-alignment, the large base resistance, $r_{b'}=2000$Ω, is the limiting factor. It gives: $t_{pd}=34.5$ps, $f_{bw} \approx f_t=$ 26.4GHz, $f_{t-}=4.42$GHz, and $f_{Max}=2.41$GHz. Using self-alignment in transistor (3), $r_{b'}$ drops by 4x (4 times) to 500Ω (still too high), t_{pd} improves by $\approx$2.5x to

20.5ps, f_{bw} (■ f_t) stays at 26.4GHz and not affected by $r_{b'}$ since base current is the input, f_{t-} increases by ∝4x to 16.2GHz, and f_{Max} increases by ∝5x to 12.9GHz. Reducing the linewidth to 0.5μm in transistor (4) decreases $r_{b'}$ further, by 2x, to 250Ω and also C_c by 40%. This improves the t_{pd} only slightly to 18.1ps from 20.5ps because the diffusion delay through the base layer, 13.7ps, dominates. f_{bw}(■f_t) is still 26.4GHz. However, f_{Max} improves by ∝2.5x to 27.6GHz from 12.9GHz, because $f_{Max}^{-2} = 16\pi^2 r_{b'} C_c \cdot (r_{b'} C_c + \omega_\alpha^{-1})$.

The examples show that as the dimension is decreased to the deep submicron range, the base doping concentration must be increased to keep the base resistance low. It also shows that the t_{pd} can be minimized and f_{t-} or f_{Max} maximized by minimizing $[r_{b'} C_{cb4} C_{eb0}/(C_{cb4}+C_{eb0})] + (K_6 X_{cb4}/\theta_{sat})$ ■ $(r_{b'} \varepsilon A_C/X_{cb4}) + (K_6 X_{cb4}/\theta_{sat})$ via varying X_{cb4} or the collector-base impurity doping N_{CC} and applied voltage V_{CB}. [K_6(large-signal) = 1, K_6(small-signal) = ½.].

The examples also show that the base diffusion delay t_B, is a significant contributor to high t_{pd}, low f_{bw}(■f_t) and f_{Max} in the submicron range (from $K_d t_B = 2.73 \times 5 = 13.7$ps in t_{pd} and 5/1.22 = 4.1ps in f_{bw}, f_t and f_{Max}). Diffusion delay during turn-on can be reduced by an aiding built-in electric field via grading the impurity concentration through the base layer. However, the built-in field slows down minority carrier extraction by the short-circuited emitter-base junction during turn-off. Thus, this improvement is not unlimited. For pulse delay, K_d drops from 2.73 to 1.71 at $\eta = 2$, and to 0.97 at $\eta = 4$. If $\eta = 4$ in the 0.5μm-BJT (4), then t_{pd} is decreased to 18.1-(2.73-0.97)x5ps = 9.3ps and f_{Max} is increased only 20%.

The examples also show that the BJT geometry and material parameter values needed to maximize the large-signal switching speed, t_{pd}^{-1}, are different than those to maximize the small-signal bandwidth, f_{bw}(■f_t), and oscillation frequency, f_{Max}.

The examples further show that f_{bw}(■f_t) is always greater than f_{t-} because $1/f_{t-}$ ∝ $r_{b'} C_c$ while f_{bw} or f_t is independent of $r_{b'}$. The error of using f_{t-}∝f_t(■f_{bw}) in f_{Max} calculation is severe in submicron BJTs. (See Fig.744.1 for illustration.)

Finally, higher electron and hole mobilities and diffusivities will shorten base-diffusion delay time, lower base resistance, and simultaneously improve the pulse delay t_{pd}, bandwidth f_{bw}(■f_t), f_{t-}, and f_{Max}. High mobility candidates are (1) GaAs whose electron mobility is up to five times higher than that in Si, (2) $Ga_xAl_yAs_z$ heterojunction BJT (HBT) which has both higher electron mobility and larger emitter energy gap, and (3) Si-emitter/Ge_xSi_{1-x}-base/Si-collector n/p/n HBT which has both higher (~twice) electron mobility and smaller energy gap in the Ge_xSi_{1-x} base. The SiGe HBT is the most promising for four reasons: (i) very high speed and (ii) high f_{Max} and (iii) tremendous beta (>10^3) at 77K due to higher mobility and smaller energy gap in the base, reported), and most importantly, (iv) easy integration into monolithic silicon integrated circuits. The HBTs will be described in sections 77n.

The numbers used for transistors (3) and (4) are typical values of the state-of-the-art Si BJT technology in 1990 which has given experimental BJTs approaching 100GHz. Si n/p/n BJTs with $t_{pd} < 10\text{ps}$ and $f_{Max} > 100\text{GHz}$ should be attainable by global optimization of the following parameters: (i) reducing X_B and grading $P_{BB} = P_{BB}(x)$ to reduce t_B, (ii) increasing $P_{BB}(x)$ to decrease $r_{b'}$, and (iii) decreasing N_{CC} of the epitaxy layer so that $(r_{b'}C_{cb}) + (X_{CB}/2\theta_{sat})$ is minimized. The formulae used to compute the values in Table 755.1 are:

$$r_{b'} = \rho_B L_{rb'}/(2L_E X_B) = L_{rb'}/(q\mu_p N_{BB}\cdot 2L_E X_B)$$

$$V_{EB-bi} = (kT/q)\log_e(N_{EE}N_{BB}/n_i^2) = 0.02585\text{x}\log_e(N_{EE}N_{BB}/10^{20})\text{V}$$

$$X_{PN4} = \sqrt{2\varepsilon_s V_B/qN_I)} = 35.96\sqrt{(10^{20}/N_I)(V_B{=}4/1\text{V})}\ \text{angstrom}$$

$$C_{pn4} = \varepsilon_s/X_{PN4} = \sqrt{q\varepsilon_s N_I/2V_B} = 28.80\sqrt{(N_I/10^{20})(1\text{V}/V_B{=}4)}\ \text{fF}/\mu\text{m}^2$$

$$1 - \alpha_{bo} = \beta_{bo}^{-1} = t_B/\tau_B$$

$$1 - \gamma_{eo} = (D_E P_E X_B/D_B N_B X_E)\exp(\Delta E_G/kT) \approx (D_E N_{BB} X_B/D_B N_{EE} X_E)\cdot 10$$

$$\beta_{fo} = 1/(1-\alpha_{bo} + 1-\gamma_{eo}) \approx 1/[10\cdot(D_E N_{BB} X_B/D_B N_{EE} X_E)+(t_B/\tau_B)]$$

$$t_{pd} = (1-\gamma_{eo})t_E + 2.73t_B + X_{CB}/\theta_{sat} + r_{b'}C_{eb0}C_{cb4}/(C_{eb0}+C_{cb4})$$

$$2\pi f_t \equiv 2\pi f_{bw} = [(1-\gamma_{eo})t_E + t_B/1.22 + X_{CB4}/2\theta_{sat}]^{-1}$$

$$2\pi f_{t-} \equiv 1/t_{t-} = [(1-\gamma_{eo})t_E + t_B/1.22 + X_{CB4}/2\theta_{sat} + r_{b'}C_{cb4}]^{-1}$$

$$f_{Max} = \sqrt{f_{t-}/8\pi r_{b'}C_{cb4}} = [4\pi\sqrt{r_{b'}C_{cb4}\cdot t_{t-}}]^{-1}.$$

Silicon energy-gap narrowing of 60meV is assumed in the emitter due to heavy doping (10^{20}cm^{-3}), which increases the minority carrier density by $\exp(60/25.85) = 10$ and makes the apparent emitter doping $N_{EE} = 10^{19}\text{cm}^{-3}$. This does not affect the switching delay and frequency response, only β_F at high currents as indicated by (738.33B), since the charge stored in the quasi-neutral emitter is small. The bias voltages of the BJT used are $V_{EB} = 0.8\text{V}$ and $V_{CB} = 3.15\text{V}$. Thus, $V_{EB-bi} - V_{EB} = 1.05-0.8 = 0.25\text{V}$, and $C_{eb0} = \sqrt{V_{EB-bi}/(V_{EB-bi}-V_{EB})}C_{eb1} \approx \sqrt{1.05/(1.05-0.8)}C_{eb1} = \sqrt{4}C_{eb1} = 2C_{eb1}$; and $V_{CB-bi} - V_{CB} = 0.86+3.15 \approx 4.0\text{V}$, and $C_{cb4} = C_{cb1}/\sqrt{4} = C_{cb1}/2$. The fastest circuits (S,CE) and (S,CB) are selected and the 90% pulse delay was computed for no build-in electric field in the base by using $\eta = 1$ in K_d to give

$$K_d = (2.06/\eta) + [(1.007/\eta)+(\eta/2.06)]^{-1} = 2.730.$$

756 Propagation Delay in Ring Oscillator

The propagation delay of a state-of-the-art BJT technology is frequently tested or bench-marked using a monolithic ring oscillator. The gate delay or propagation delay per transistor is approximately given by the oscillation period divided by the number of transistors in the ring.

In the ring oscillator circuit, N identical transistors (with monolithic load resistors) are connected in cascade in a ring or a closed loop. Consider the CE BJT inverters shown in Fig.756.1(a). Each is characterized by a dependent current source, $g_m v_n$, an input admittance, Y_i, and an output admittance Y_o, which also include all the parasitic admittances from the transistors and the inter-transistor connection lines. To make the circuit oscillate, an odd number of BJTs are connected in a ring shown in Fig.756.1(b). It will oscillate when forward base-emitter and reverse collector-base voltages are applied to all BJTs. It is evident that this circuit has a Fan-In and Fan-Out of 1.

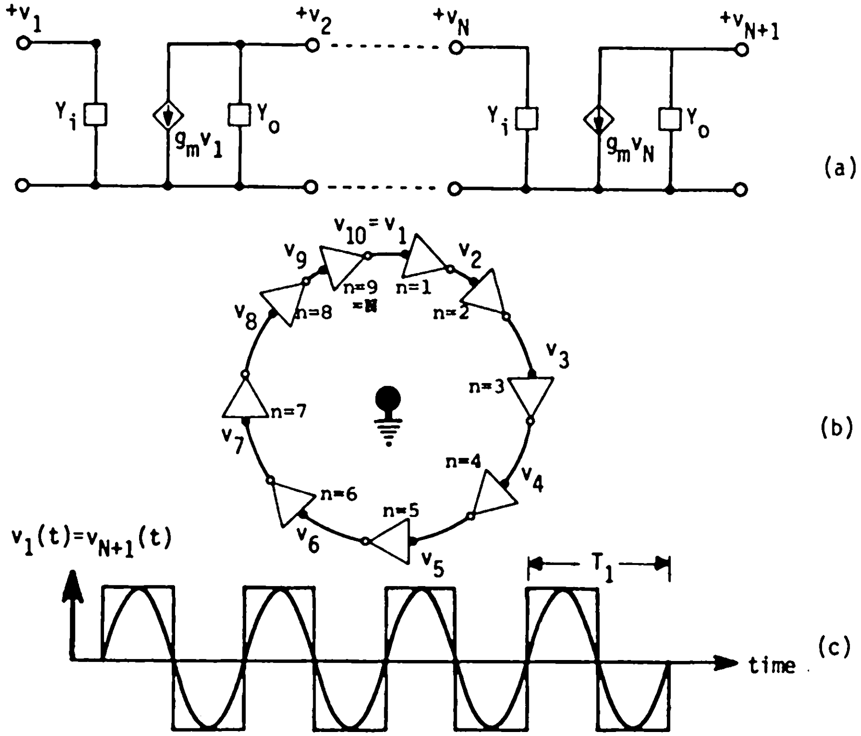


Fig.756.1 Ring oscillator of CE BJTs. (a) The simplified equivalent circuit. (b) The ring circuit topology. The triangle is the nonlinear large-signal CE BJT inverter-amplifier. (c) The voltage waveform at input and output node.

The voltage waveform of the oscillation is shown in Fig.756.1(c) whose period is designated by T_1. In principle, it should be nearly a square wave, however, alpha drop-off at higher frequencies and circuit capacitance and inductance in the ring will filter out the higher harmonics and restrict the oscillation to the fundamental frequency shown by the sinusoidal wave in Fig.756.1(c). The voltage amplitude of the oscillation is limited by the power supply voltage.

In order to use T_1 as a figure-of-merit, a formula relating T_1 to the transistor parameters is required. To derive this formula using a minimum amount of mathematics, we shall make a single-frequency analysis of this nonlinear problem by treating the BJT as linear at the frequency $\omega_1 = 2\pi f_1 = 1/T_1$ in order to be able to use the BJT small-signal equivalent circuit in the analysis. Thus, let us denote the ring oscillator characteristic admittance, Y, by

$$Y = Y_i + Y_o = (G_i+G_o) + j\omega(C_i+C_o) \equiv G + j\omega C. \qquad (756.1)$$

Then the circuit in Fig.756.1(b) gives

$$\begin{aligned} v_{N+1} &= -(g_m/Y)v_N = +(g_m/Y)^2 v_{N-1} = \cdots\cdots = -(g_m/Y)_N v_1 \\ &= v_1 \qquad (N=\text{odd}) \qquad (756.2) \end{aligned}$$

where N=odd. N=even will not oscillate because the output and the input are in phase (see Problem P756.1). Using the expansion

$$\begin{aligned} (-1)^{1/N} &= \exp[j(2k-1)\pi/N] \qquad (k=1,2,3,\ldots.) \\ &= \cos[(2k-1)\pi/N] + j\sin[(2k-1)\pi/N] \qquad (756.3) \end{aligned}$$

then

$$\begin{aligned} Y/g_m &= G/g_m + j\omega C/g_m = (-1)^{1/N} \\ &= \cos[(2k-1)\pi/N] + j\sin[(2k-1)\pi/N] \qquad (756.4) \end{aligned}$$

so

$$G/g_m = \cos[(2k-1)\pi/N] \qquad (756.5A)$$

and

$$\omega C/g_m = \sin[(2k-1)\pi/N]. \qquad (756.5B)$$

To satisfy these simultaneously, we must have

$$\omega_k C/G = \tan[(2k-1)\pi/N] \qquad (k=1,2,3,\ldots.) \qquad (756.6)$$

which determines the condition of oscillation, i.e. the oscillation frequency, ω_k. For the lowest frequency or the fundamental component, k=1, the oscillation frequency is

$$\begin{aligned} \omega_1 C/G &= 2\pi f_1(C/G) = (2\pi/T_1)(C/G) \qquad (756.7) \\ &= \tan[\pi/N] \simeq \pi/N \qquad \text{if } N \geq 11. \qquad (756.7A) \end{aligned}$$

Thus, the **ring-oscillator average** (50% of pull-up + pull-down) **delay per transistor** is

$$T_1/2N = C/G. \tag{756.8}$$

The capacitance and conductance can be related to the BJT internal and parasitic parameters. An inspection of the CEss-Hπ model of Fig.743.1(d) shows that if (i) $r_{e'}$ and $r_{c'}$ can be neglected, (ii) g_o, g_μ and g_π are small compared with g_{mo}, and (iii) C_{cs} due to the p/n junction well to isolate the collectors of the N BJTs is added, then

$$\begin{aligned} C &= C_i + C_o \\ &= (C_{ul}+C_{eb}+C_c) + (C_{ul}+C_{cs}+C_c) = 2(C_{ul}+C_c)+C_{eb}+C_{cs} \quad (756.9A) \end{aligned}$$

and

$$\begin{aligned} G &= G_i + G_o \\ &\simeq g_m + 0 = g_{mo}/(1+j\omega/\omega_{gm}) \simeq g_{mo}. \quad (756.9B) \end{aligned}$$

The low-frequency approximation for g_m is valid since the oscillation frequency $\omega=\omega_1(=\pi G/CN) << \omega_{gm}$ when N is large. Thus, the above results of C and G can be used in (756.8) to give an **average ring-oscillator gate delay** of

$$T_1/2N = C/G = [2(C_{ul}+C_c)+C_{eb}+C_{cs}]/g_{mo}. \tag{756.10}$$

Using the numerical result of the four transistors in Table 755.1 and assuming $g_{mo}=1mS$ at $I_C=25\mu A$, $C_{ul}=C_c$, and using

$$C_{eb} = C_e+C_{ebt}+C_b \simeq C_{ebt}+C_b = C_{ebt} + g_{mo}t_B,$$

then the average ring oscillator gate delay per transistor at $V_{CB\text{-}bi}-V_{CB}=4V$ and $V_{EB\text{-}bi}-V_{EB}=0.25V$ for the four transistors in Table 755.1 are respectively given by

$$\begin{aligned} T_1/2N &= [2(30.1x2/2)+2x10.2+500+19.2/2)fs]/10^{-3} = 590ps && (756.11) \\ &= [2(30.1x2/2)+2x10.2+5 \;\; +19.2/2)fs]/10^{-3} = \;\; 95ps && (756.12) \\ &= [2(15.4x2/2)+2x10.2+5 \;\; + 6.8/2)fs]/10^{-3} = \;\; 60ps && (756.13) \\ &= [2(9.24x2/2)+2x5.1 +5 \;\; + 3.4/2)fs]/10^{-3} = \;\; 35ps. && (756.14) \end{aligned}$$

These numerical results show that for the 1-μm non-selfaligned-emitter transistor, BJT-(1) in Table 755.1, the ring-oscillator gate delay is limited by base diffusion delay, $t_B=500ps$. For the three 0.1-μm base transistors, self-alignment reduces the ring-oscillator gate delay from 95ps to 60ps via reduction of parasitic areas so the capacitances are smaller. When the base width is reduced to 0.5μm, (756.14), capacitances are reduced further via reduction of parasitic areas of the collector-substrate isolation junction and underlap collector-base junction. It is evident that ring oscillation is easy to measure since the oscillation frequency is low. For example, at N=21, the fastest BJT, transistor (4), would oscillate with a period of $T_1=2x21x35ps=1.5ns$ or a frequency of $10^9/1.5=680MHz$.

760 CIRCUIT APPLICATIONS OF BIPOLAR JUNCTION TRANSISTOR

Application examples of BJT in digital switching circuits will be described in the following sections, 76n. Like that of MOSFET given in sections 67n, only the simplest basic circuit building blocks which contain one or two cascaded BJTs are described. The scope is further limited to those that have found great success in volume applications at the market place. Pedagogical rather than chronological description of the evolution of the circuits will be emphasized. Although there are many interesting and important BJT analog circuits which perform crucial functions for a system, they are not described here. Many important analog functions are now performed by digitizing the analog signal, for example, the digital audio recorder and digital high definition television, further increasing the market share of digital integrated circuits. The basic digital circuits to be described are selected to serve as an introduction and connection to the more complex many-BJT and multi-stage or multi-gate digital integrated circuits given in intermediate and advanced digital transistor circuit textbooks. The descriptions of circuit operation to be given emphasize device physics in order to supplement the circuit textbooks which focus on circuit analysis and often do not give adequate device physics.

761 The BJT Digital Inverters

The inverter circuit is the fundamental building block of digital logic circuits and networks. Its two output voltage states, positive and negative or zero (HIGH and LOW), are needed to represent the binary numbers 1 and 0 in all digital logic circuits and networks. Only the collector output of the CE configuration of BJT is the inversion of the input. The emitter output of the CE and both the collector and base outputs of the CB configurations are not inverted from the input. These are illustrated in circuit diagrams of an n/p/n BJT in Figs.761.1(a) for CB and (b) for CE. The collector d.c. supply voltage is $V_{CC} = +5V$. The two configurations contain both the source resistance (R_E or R_B) and load resistance (R_C and R_B or R_E) to illustrate the output/input voltage inversion. In practice, these source and load resistances are used in multistage cascaded circuits to give the proper d.c. bias and to match the impedance (50-Ω) of the interconnect monolithic transmission lines on a chip. In lower speed circuits of the earliest (~1965) generation bipolar technology, these resistances were quite large ($\propto > 10$k) compared with the internal base, source, and collector resistances so that the effects from production variations of the internal resistances are diminished. The effects of these external resistances on performance are similar to those of the source and load capacitances in MOS digital circuits described in section 673. In MOS digital circuits, the MOST only needs to charge and discharge capacitances. However, in bipolar digital circuits, the BJT not only needs to charge and discharge the load and circuit capacitances but also drives the source and load resistances. The intrinsic delays (drift transit time of minority carriers in MOSFET inversion channel and diffusion transit time of minority carriers in BJT base) do not control their switching speeds because of large C load in MOS circuits, and large C and possibly small R loads in BJT circuits. To attain the maximum switching speed or frequency of operation, the RC delays must be reduced so much that base diffusion transit time as well as collector and emitter transit times will become the speed limiting factors. The analyses in sections 751-753, and especially the comparison in 754 and numerical examples in 755, have considered these BJT intrinsic delays in the limit of zero external resistance loading. These zero loading results can be advantageously used to analyze and design speedy BJT inverters which we shall describe in the following paragraphs of this section. Specific examples are given in the following three sections.

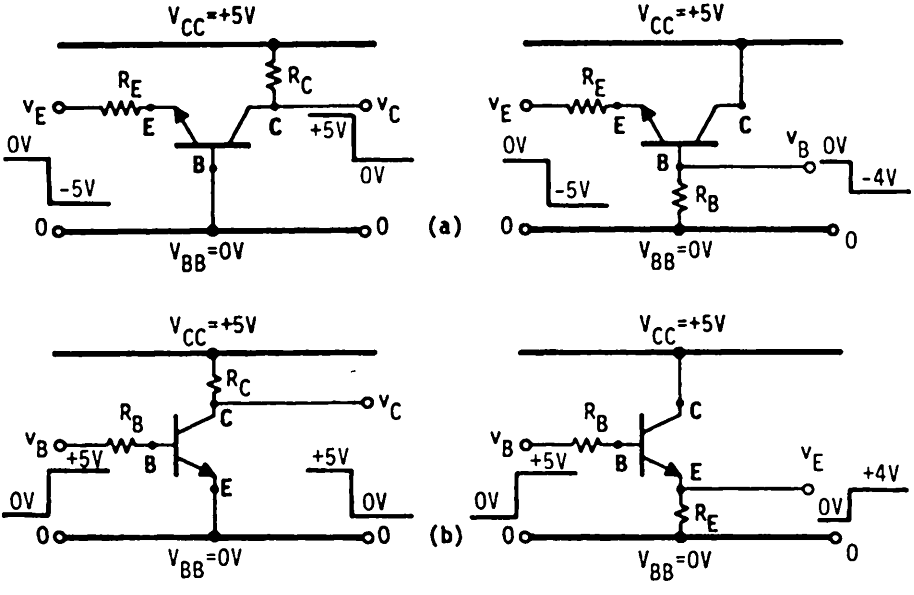


Fig.761.1 The two one-transistor building blocks for digital BJT integrated circuits with inverted or non-inverted (follower) outputs. (a) The CB configuration. (b) The CE configuration.

An inspection of the CB circuits in Fig.761.1(a) readily shows that both the collector and base output voltages, $v_C(t)$ and $v_B(t)$, are in phase with the emitter input voltage, $v_E(t)$. For example, if $v_E(t)$ is 0V and decreases towards -5V, then $v_C(t)$ at +5V also decreases but towards $V_{CB\text{-sat}}<0V$, and $v_B(t)$ at 0V also decreases towards $-5V+V_{EB\text{-bi}}$. These outputs are known as the **collector-follower** and **base-follower** outputs.

A similar examination of the CE circuits in Fig.761.1(b) shows that output $v_C(t)$ is the **inversion** of input $v_B(t)$, while output $v_E(t)$ follows input $v_B(t)$ and is known as the **emitter-follower** output.

Although the non-inverted CB collector and base outputs and the non-inverted CE emitter output cannot be used alone to give the two logic states, each can be connected in cascade with the inverted CE collector output to make a 2-transistor inverter building block. The input BJT is either in the CB collector-output configuration (less than unity current gain) or the CE emitter-follower configuration (less than unity voltage gain). The output BJT is in the CE collector-output configuration, not only to invert the signal but also to make up for the attenuation. Thus, both the current and voltage amplifications of the 2-transistor inverters can be greater than unity which is necessary to cascade many inverters in order to give desired logic functions. (See example in Fig.765.1.) The CB input BJT also speeds up the inverter. During turn-on, it gives a fast collector output current with an intrinsic average propagation delay of $0.76t_B$ from (754.6) and a 90% pulse delay of $2.73t_B$ from (755.7). The fast CB output current can drive a monolithic resistance to give a fast voltage pulse to drive the CE output stage. This speeds up the CE stage to an average propagation delay of $0.76t_B$ and a 90% pulse delay of $2.73t_B$ prior to saturation. The CB input BJT also speeds up the turn-off transient because when its emitter is shorted to ground, its turn-off collector-to-emitter current helps to pull out the stored base charge in the output BJT.

762 The Common-Emitter BJT Inverter

The one-transistor CE BJT inverter circuit in Fig.762.1(a) will be analyzed since it is the basic building block. In contrast to many recent textbooks but following many older textbooks, the p/n/p will be described. This choice gives a positive base minority carrier (holes) charge which avoids double negative conversion during verbal or mental deliberations of the device physics underlying the output waveforms for a range of circuit element values, but at the expense of negative (rather than positive) input and output voltage waveforms. In practice, n/p/n is used due to higher speed and easier fabrication. This analysis can be readily applied to n/p/n by a mere change of sign of the voltages and currents.

The emitter resistance, R_E, is removed in this analysis and it will be added later to describe the 2-transistor inverters using the emitter follower. The base and collector resistances are $R_B = 10k\Omega$ and $R_C = 1k\Omega$. $\beta_F = 100$ is assumed so that $\beta_F >> (R_B/R_C) = 10k/1k = 10$ in order to allow for fan-outs up to 10. A power supply of $V_{CC} = -5V$ is assumed. The static loadline and quasi-static locus on the I_C-V_{CB} plane are shown in Fig.762.1(b) which is the key aid to provide a mental image of the (current,voltage) transient because it indicates the initial and final states on a familiar BJT characteristic diagram. It also shows the true loci of the transient if the transient is slow. The input and output voltage waveforms are shown in Figs.762.1(c) and (d) and the excess hole concentration in Fig.762.1(e).

The output collector current or voltage waveform shown in Fig.762.1(d) has six phases, three each for the turn-on and turn-off transients. The physical origin, approximate analytical solution, and numerical examples of the six phases are given in the following six subsections. More detailed solutions of some of the phases are given at the end of this section in order to focus on the fundamentals first.

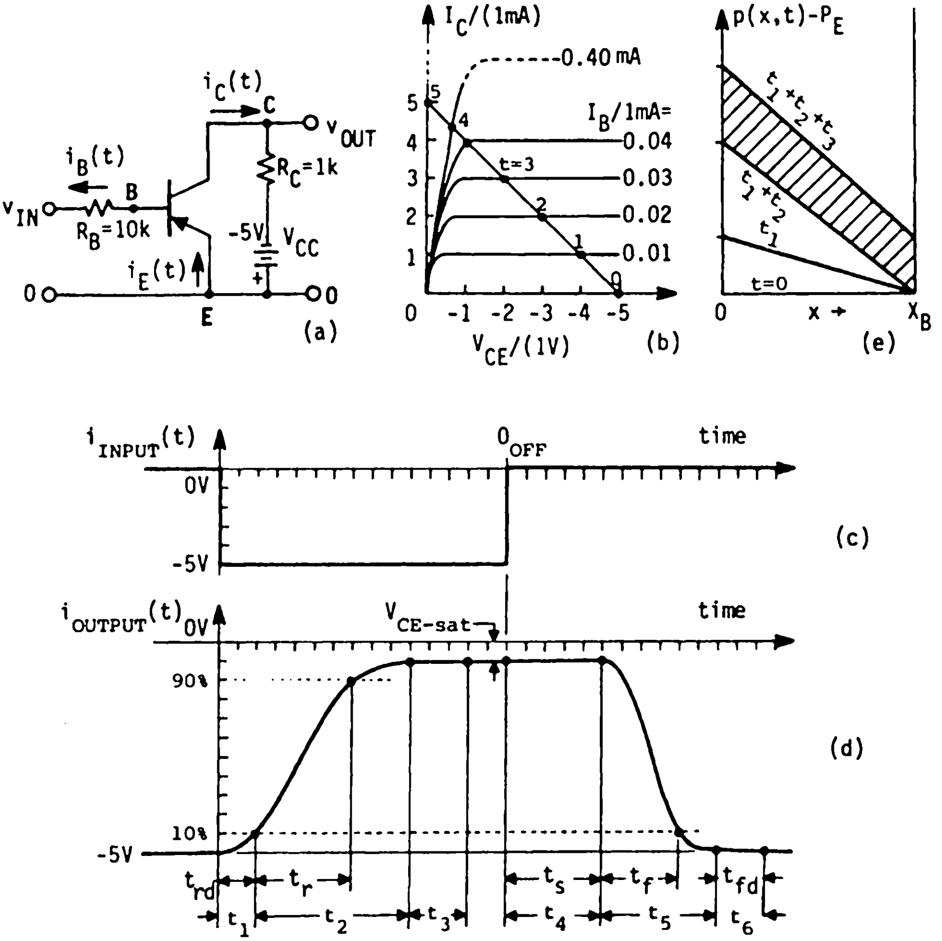

$i_C(t)$
C
V_{OUT}
$i_B(t)$
B
V_{IN}
$R_B=10k$
$R_C=1k$
-5V
V_{CC}
$i_E(t)$
0
0
E
(a)
$I_C/(1mA)$
0.40 mA
$I_B/1mA=$
0.04
0.03
0.02
0.01
t=3
5
4
3
2
1
0
0 -1 -2 -3 -4 -5
$V_{CE}/(1V)$
(b)
$p(x,t)-P_E$
$t_1+t_2+t_3$
t_1+t_2
t_1
t=0
0
x →
X_B
(e)
$i_{INPUT}(t)$
0_{OFF}
time
0V
-5V
(c)
$i_{OUTPUT}(t)$
0V
V_{CE-sat}
time
90%
10%
-5V
(d)
t_{rd}
t_r
t_s
t_f
t_{fd}
t_1
t_2
t_3
t_4
t_5
t_6

Fig.762.1 The circuit diagram and waveforms of a practical common-emitter p/n/p BJT voltage inverter. (a) The circuit diagram. (b) The static output characteristics and the quasi-static locus during the turn-on and turn-off transient. (c) The input base voltage waveform. (d) The output collector voltage waveform. (e) The space-time variation of the excess minority carrier (hole) concentration in the quasi-neutral base layer.

(1) Charging Up the Emitter-Base Space-Charge Layer Capacitance

Phase 1 ($t=0$ to t_1) is the initial delay due to charging up the base-emitter n/p junction space-charge layer capacitance, C_{be}, by the majority carriers or base current (electron current). The delay is designated by t_{rd} ($=t_1$) known as the initial rise delay time and its value is similar to the phase I delay in the CB configuration analyzed in (752.6) to (752.22) but $r_{b'}$ is replaced by the total base resistance (internal plus external, $r_{b'}+R_B$). Thus, from (752.22) $t_1 = t_{dr} = (R_B+r_{b'})C_{be0} =$

$(10k+1.25k)13.09fF = 147ps \approx 0.15ns$. From (752.11A), $V_{EB'}/V_{EB\text{-}bi}=0.625$, then the base-emitter voltage at t_1 is $v_{EB'}(t_1)=0.625(kT/q)\log_e 10^{16} = 0.595V$.

(2) Charging Up the Quasi-Neutral Base in the Active Range

Phase 2 is the duration of $t=t_1$ to t_1+t_2. As we enter this phase at t_1, the forward emitter-base junction voltage, $v_{EB'}$, begins to exceed the injection threshold and many minority carriers (holes) are injected from the p-emitter into the n-base. These injected carriers reach the collector, increases the collector current, increases the voltage drop across the collector load resistance R_C, and decreases the magnitude of the output voltage, $v_O(t)$, causing it to rise from -5V towards 0V as shown in Fig.762.1(d). The time from 10% to 90% of final collector voltage is known as the **rise time**, t_r. The locus of this phase lies in the active region and is shown in Fig.762.1(b).

The collector current transient waveform was analyzed in section 753 using the charge-control method. The results were given by (753.4A) and (753.5). They can be used here because the reverse base charge is negligible during phase 2, i.e., $q_{BR}(t<t_1+t_2)=0$, due to large instantaneous reverse collector-base junction voltage from the large applied reverse collector bias, $V_{CC}=-5V$. Thus, using $q_{FB}(t)$ from (753.4A) and $q_B(t)=q_{BF}(t)+q_{BR}(t)=q_{FB}(t)+0$, then

$$q_B(t) = q_{BF}(t) + q_{BR}(t) \approx q_{BF}(t) = J_B\tau_{BF}[1 - \exp(-t/\tau_{BF})] \qquad (762.1)$$

$$j_C(t) = q_B(t)/t_{BF} \approx q_{FB}(t)/t_{BF} = \beta_F J_B[1 - \exp(-t/\tau_{BF})] \qquad (762.2)$$

$$i_C(t) = J_C A_E = \beta_F I_B[1 - \exp(-t/\tau_{BF})] \qquad (762.3)$$

$$i_B(t) \approx dq_{BF}(t)/dt + q_{BF}(t)/\tau_{BF} = I_B[1 - \exp(-t/\tau_{BF})]. \qquad (762.4)$$

Appropriate junction area is to be multiplied to give the total current and charge. Note that the time constant of phase 2 is the minority carrier recombination lifetime in the base, τ_{BF}. From the circuit, Fig.762.1(a), the maximum or steady-state base current is given by

$$I_B = -(V_{BB}-V_{EB'})/(R_B+r_{b'}) < (5-0.595)/(10k+1.25k) = 0.39mA \qquad (762.4A)$$

where we have used the value of $v_{EB'}(t_1)=0.595V$ at the start of phase 2 instead of the higher end value, $v_{EB'}(t_1+t_2)$ of 0.8 to 0.9V.

The collector current of phase 2, given by (762.2) or (762.3), cannot continue to conclusion, for example, reaching 90% of its final value $i_C(t=t_{0.9}) = 0.9\beta_F I_B$. For $\beta_F=100$, this is $0.9\beta_F x0.39mA = 0.35\beta_F mA = 35mA$. The reason is that this current is seven times higher than the maximum collector current that can be drawn through the collector load resistance $|V_{CC}|/R_C = |-5|/1k = 5mA$, which is also indicated by the intersection of the load line with the current axis, labeled $t=5$ in Fig.762.1(b). Voltage drop across $r_{e'}$ and $r_{c'}$ and across the CB and EB junctions will reduce this by about 1mA to 4mA as indicated by the intersection of the load line with the saturation line, labeled $t=4=t_1+t_2$ in Fig.762.1(b). At $t=4=t_1+t_2$, the BJT enters the saturation region from the active region; phase (2) ends and phase (3) begins. The current and voltage are

$$i_C(t=t_1+t_2) = -(V_{CC}-V_{CE\text{-}sat})/R_C \simeq I_{C\text{-}sat} \simeq I_C \qquad (762.5A)$$

and

$$v_O(t=t_1+t_2) = V_{CE\text{-}sat}. \qquad (762.5B)$$

The CE saturation voltage, $V_{CE\text{-}sat}$, was given by (735.41A). Substituting these into (762.3), the length of phase 2 and the 10% to 90% output-voltage rise time are

$$t_2 = \tau_B\cdot\log_e\{1 - (I_{C\text{-}sat}/\beta_F I_B)\}^{-1} \qquad (762.6)$$

$$\simeq \tau_B\cdot I_{C\text{-}sat}/(\beta_F I_B) = (I_{C\text{-}sat}/I_B)t_B \qquad (762.6A)$$

$$t_r = \tau_B\cdot\log_e\{[1-0.1(I_{C\text{-}sat}/\beta_F I_B)]/[1-0.9(I_{C\text{-}sat}/\beta_F I_B]\} \qquad (762.7)$$

$$\approx 0.8(I_{C\text{-}sat}/I_B)t_B = 0.8t_2. \qquad (762.7A)$$

Using the $t_2 << \tau_B$ approximation, the total base charge from (762.1) is

$$q_B(t_2) \approx I_B t_2 = I_{C\text{-}sat}\cdot t_B. \qquad (762.7B)$$

For the 0.1μm base-thickness Si p/n/p transistor,

$$t_B = X_B^2/2D_B = (10^{-5})^2/(2x0.02585x250) = 7.74ps \qquad (762.8A)$$

and

$$q_B(t_2) = I_{C\text{-}sat}\cdot t_B = (5-1)mAx7.74ps = 38.7fC \qquad (762.8B)$$

$$= (38.7x10^{-15}/1.6x19^{-19})q = 0.24x10^6q \simeq 240kq$$

$$\approx 0.25 \text{ million holes.} \qquad (762.8C)$$

These are increased by 100x (100 times) to 774ps and 3.87pC or 25 million holes if the base is 10 times thicker, $X_B=1\mu$m. The large number of holes is precisely the reason that the 1-μm BJT can drive a much larger capacitance load than the 1-μm MOSFET whose inversion channel contains much fewer electrons or holes (only about 1M). At 0.1-μm, the FET and BJT contains essentially the same number of signal holes, then, the BJT is not a much faster driver. The duration of phase 2 for the 0.1μm base p/n/p BJT is then

$$t_2 = (I_{C\text{-}sat}/I_B)\cdot t_B$$

$$\approx [5mA/0.39mA]*7.74ps \approx 100ps \quad \text{for } X_B=0.1\mu m \qquad (762.8D)$$

and

$$\approx [5mA/0.39mA]*774ps \approx 2\to 10 \text{ ns} \quad \text{for } X_B=1.0\mu m. \qquad (762.8E)$$

(3) Further Charging Up the Quasi-Neutral Base in the Saturation Range

Phase 3 spans the time $t=t_1+t_2$ to $t_1+t_2+t_3$ with duration t_3. Phase 2 ends at $t=t_1+t_2$ and phase 3 begins when the collector current $i_C(t)$ and collector-emitter terminal voltage $v_{CE}(t)$ reach the saturation values $I_{C\text{-}sat}$ and $V_{CE\text{-}sat}$. Even though they no longer change with time appreciably after the end of phase 2, the base current continues to rise from the value at t_1+t_2

$$i_B(t=t_1+t_2) = I_{C\text{-}sat}/\beta_F = 4mA/100 = 0.04mA \qquad (762.9A)$$

to the final steady-state value given by (762.4A)

$$i_B(t=t_1+t_2+t_3) = -(V_{BB} - V_{BE\text{-}sat})/(R_B+r_{b'}) \qquad (762.4A)$$

$$= -[-5-(-0.8)]/(10k+1.25k) = 0.373\text{mA}. \qquad (762.9B)$$

This increases the emitter-base junction forward bias, $v_{EB'}(t)$, from 0.595V to 0.8V (or 0.9V depending on τ_B), which injects more holes into the base from the emitter. The increasing base current will also increase the collector current slightly to cause the already small base-collector voltage,

$$v_{CB'}(t=t_1+t_2) = v_{CE} - v_{EB'} = -1\text{V} - 0.595 = -1.6\text{V} \qquad (762.10A)$$

to decrease, cross zero, and become **forward biased**. This may be visualized from the circuit diagram given in Fig.762.1(a) by letting $i_B(t)$ increase with time from 0.04mA to 0.373mA and $i_C(t)$ increase from 4mA to 4.3mA. The voltage equation of the collector-emitter loop is

$$i_B(t)(R_B+r_{b'}) - 5\text{V} = v_{B'C}(t) + i_C(t)R_C - 5\text{V}$$

or

$$v_{CB'}(t) = i_C(t)\cdot R_C - i_B(t)\cdot(R_B+r_{b'}) \qquad (762.10B)$$

$$= -4.3\text{mA}\times 10k - 0.373\times(11.25k) = +0.104\text{V}. \qquad (762.10C)$$

At the end of phase 3, $t=t_1+t_2+t_3$,

$$v_{CB'}(t_1+t_2+t_3) = V_{CE\text{-}sat} - V_{EB'\text{-}sat}$$

$$= -0.1 - (-0.8) = +0.7\text{V}. \qquad (762.10D)$$

This gives the origin and the sequence of events leading to the forward bias on the collector-base junction. Thus, the base charge (holes), $q_B(t)$, will increase during phase 3 from two sources: (1) the further increase of the forward voltage on the emitter-base junction giving more $q_{BF}(t)$, and (2) the collector-base junction voltage changing from reverse to forward bias which injects some minority carriers into the base from the collector, giving an increasing $q_{BR}(t)$. The space-time dependence of $p(x,t)-P_B$ from these two sources is illustrated in Fig.762.1(e). The neglected reverse base charge, $q_{BR}(t)$, must now be included in the charge-control analysis of this phase.

The charge-control equation of the base charge now contains two components in the saturation range. From (751.8B) or (751.13A), it is

$$j_B(t) = dq_{BF}/dt + q_{BF}/\tau_{BF} + dq_{BR}/dt + q_{BR}/\tau_{BR} \qquad (751.13A)$$

$$= dq_B/dt + q_B/\tau_B \qquad (762.11)$$

where by definition, $q_B = q_{BF}+q_{BR}$, and $\tau_{BF} = \tau_{BR} = \tau_B$. The equality given by $\tau_{BF}=\tau_{BF}=\tau_B$ is physically important because there is only one lifetime for the minority carriers (holes) in the base layer regardless of its location of injection, whether from the p-emitter or p-collector, unless the density of the recombination centers is spatially varying through the very thin base layer which is unlikely.

Nevertheless, two empirical lifetimes, one for forward and one for reverse, have been used by SEEC and the subsequent textbooks that follow SEEC, to represent surface recombination and underlap collector-base junction losses. A physically correct and more accurate representation is by partition of the BJT and addition of the parasitic surface recombination, underlap diode, and series resistances to the intrinsic transistor, instead of defining physically unrealistic empirical lifetimes. Indeed (762.11) follows rigorously from the original charge-control equation, (751.8B), and gives the correct result for well-designed high-performance BJTs. Thus, the total base charge continues to increase with the time constant τ_B as indicated by (762.11) to the final steady-state value given by $I_B\tau_B$. The collector current saturates to $I_{CE\text{-sat}}$ and the collector-emitter voltage saturates to $V_{CE\text{-sat}}$. The time to reach 10% and 90% of the final base charge are simply $t_{0.1}=\tau_B\log_e(1/0.9)=0.105\tau_B$ and $t_{0.9}=\tau_B\log_e 10=2.303\tau_B$ and the 10% to 90% rise time is $t_{0.9\text{-}0.1}=(2.303\text{-}0.105)\tau_B=2.198\tau_B\propto 2.20\tau_B$.

There is a second transient component in $i_B(t)$ which is β_P-times smaller and $2\beta_F$-times faster and usually neglected compared with the main τ_B transient. Its time constant is $t_{FAST} = t_B/2 = X_B{}^2/4D_B$ or $t_B/(\eta+\eta^{-1})$ with field acceleration $(\eta>1)$. It originates from the diffusion delay of the minority carriers (holes), injected by the forward-biased collector-base junction and passing through the base layer to reach the emitter-base junction. It is just like the initial delay in $i_C(t)$ of the CB configuration analyzed in section 752.

A third transient in $i_B(t)$ is also small and fast. It comes from charging the collector-base junction space-charge layer capacitance by the majority carriers through the base resistance. It is similar to that of the CB configuration described in section 752. It can again be neglected compared with the main transient with time constant τ_B even when the total base resistance is increased from $r_{b'}=1.25k$ to $r_{b'}+R_B=11.25k$ because $\tau_B\propto 1\mu s>>(r_{b'}+R_B)C_{cb}\lesssim 1ns$ for $C_{cb}=0.1pF$.

(4) Discharging the Stored Base Charge in the Saturation Range

The turn-off transient begins when the input base voltage is switched from -5V to 0V at $t=0_{OFF}$ as indicated in Fig.762.1(c). The collector current and output voltage can be divided into three phases as indicated in Fig.762.1(d). The BJT is in the saturation region during the first turn-off phase (phase 4) with t_4 in Fig.762.1(d). The collector-base voltage changes from the initial forward value to zero. At zero bias, $q_{BR}(t)=0$ and stays 0, and the BJT returns to the active region.

The base charge transient is governed by (762.11) whose time constant is the minority carrier recombination lifetime in the base, τ_B. During the final turn-on phase when the BJT was in saturation, the base charge was decomposed into two components shown in Fig.762.1(e) and given by

$$q_B(t) = q_{BF}(t) + q_{FR}(t) \qquad \{\text{Turn-on saturation range}\} \qquad (762.12A)$$

These two components are delineated in Fig.762.2(a). During the turn-off, the base charge can also be decomposed, but into two different components from those during turn-on. The turn-off decomposition in the saturation range, shown in Fig.762.2(b), is

$$q_B(t) = q_B(0) + q_S(t) \qquad \{\text{Turn-off saturation range}\} \qquad (762.12B)$$

which is valid not only in the turn-off but also turn-on saturation range. Here, $q_S(t)$ is the time dependent component in the saturation region shown in Fig.762.2(b). This component extends over both the previous phase 3 and the present phase 4. The initial base charge, $q_B(0)$, is the constant component during the saturation phases (phases 3 and 4). It is the charge stored in the base during phase 2 before reaching saturation at the beginning of phase 3 and given by (762.7B),

$$q_B(0) = q_B(t_1+t_2) = I_B t_r = I_{C\text{-sat}} \cdot t_B \equiv I_C \tau_B/\beta_F \qquad (762.13)$$

where I_C is understood to be the I_C at saturation for the given base drive current I_B, collector supply voltage V_{CC}, and load resistance R_C.

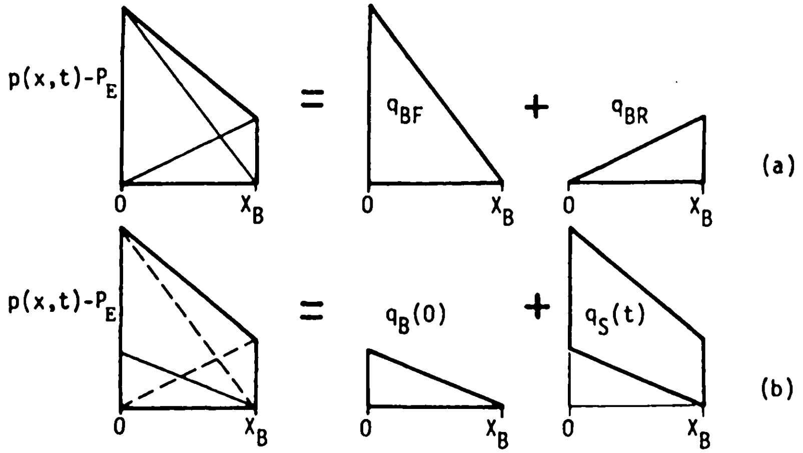


Fig.762.2 The two alternative decompositions of the space-time dependence of the base minority carrier (hole) charge concentration $q_B(t)$. (a) Turn-on range only, $q_B(t) = q_{BF}(t) + q_{BR}(t)$. (b) Both turn-on and turn-off ranges, $q_B(t) = q_B(0) + q_S(t)$.

In order to simplify the notation, we shift the $t=0$ to the end of phase 2 or $t=t_1+t_2$, labeled in Fig.762.1(d), which is implied in (762.12B). Then, using (762.12B), the base current equation, (762.11), can be re-written as

$$j_B(t) = dq_S(t)/dt + q_B(0)/\tau_B + q_S(t)/\tau_B$$

which can be rearranged using (762.13) for $q_B(0)$ to give

$$i_B(t) - I_C\tau_B/\beta_F = dq_S(t)/dt + q_S(t)/\tau_B \qquad (762.14)$$

where q_S is the total charge rather than areal charge density.

Equation (762.14) is the general equation in the saturation range. It can be applied to both the turn-on (phase 3) and turn-off (phase 4) saturation ranges for a variety of base current drives to speed up the BJT switch:

$$i_B(t) = |[V_{BB}-v_{BE}(t)]|/(R_B+r_{b'})$$

$$\approx |I_{B1}| > |I_C/\beta_F| \qquad \{\text{Phase 3. Overdrive turn-on.}\} \qquad (762.15)$$

$$= |-v_{BE}(t)|/(R_B+r_{b'}) \approx 0 \qquad \{\text{Phase 4. Free turn-off.}\} \qquad (762.16)$$

$$= -|I_{B2}| \qquad \{\text{Phase 4. Overdrive turn-off.}\} \qquad (762.17)$$

The condition of base current exceeding the maximum collector current, $I_{B1} > I_C/\beta_F$ given by (762.15) for phase 3, is known as **overdrive turn-on** and the ratio, $I_{B1}/(I_C/\beta_F)$, is known as the **turn-on overdrive factor**. This ratio determines the amount of saturation charge stored in the base from the forward-biased collector-base junction that must be removed to turn off the transistor. Thus, the turn-off storage delay time, t_s, or the duration of phase 4 is directly related to the turn-on overdrive factor which will be calculated in the following paragraph. The free turn-off condition given by (762.16) with the voltage source shorted is approximately the same as that with the voltage source opened since the base input resistance, R_B, is large. The phase 4 turn-off storage delay, t_s, can be shortened by applying a reverse base current, $-|I_{B2}|$, to draw out the stored holes as indicated by (762.17). This is known as **overdrive turn-off** and the **turn-off overdrive factor** is $|I_{B2}|/(I_C/\beta_F)$.

The phase 4 duration, t_s, is due to extracting the stored hole injected by both the emitter-base and collector-base junctions when the collector-base junction is forward biased. The storage time from the solution of (762.14) is

$$q_S(t) = (I_{B1}-I_C/\beta_F)\tau_B - (I_{B1}+I_{B2})\tau_B[1 - \exp(-t/\tau_B)] \qquad (762.18)$$

where $I_{B2} > 0$ and flowing in the reverse direction of the base-emitter n/p junction diode. The duration of phase 4 or the turn-off **storage delay time** is obtained from $q_S(t=t_s)=0$, giving $t_4=t_s$ which is

$$t_s = \tau_B \log_e[I_{B2}+I_{B1}]/[I_{B2}+(I_C/\beta_F)] \qquad \{\text{Overdrive turn-off}\} \qquad (762.19)$$

$$= \tau_B \log_e(\beta_F I_{B1}/I_C) \qquad \{\text{Free turn-off}\} \qquad (762.19A)$$

$$\approx \tau_B \log_e[1+(I_{B1}/I_{B2})] \qquad \{\text{Large overdrives or high } \beta_F\}. \qquad (762.19B)$$

Equation (762.19B) shows that the storage time can be reduced to nearly zero (or $<< \tau_B$) by a large reverse base current, $I_{B2} >> I_{B1}$. This is illustrated by the collector voltage waveform in Fig.762.3(b). This turn-off speed-up via overdrive is identical to the turn-off speed-up in a p/n junction diode given by (553.3) where $J_F = I_{B1}$ and $J_R = I_{B2}$ is the overdrive. However, this may not be a practical method because a second power supply is needed to supply the reverse base current while integrated circuits are normally designed to use only a single 5V power supply. Two other methods to shorten the turn-off storage delay are discussed later.

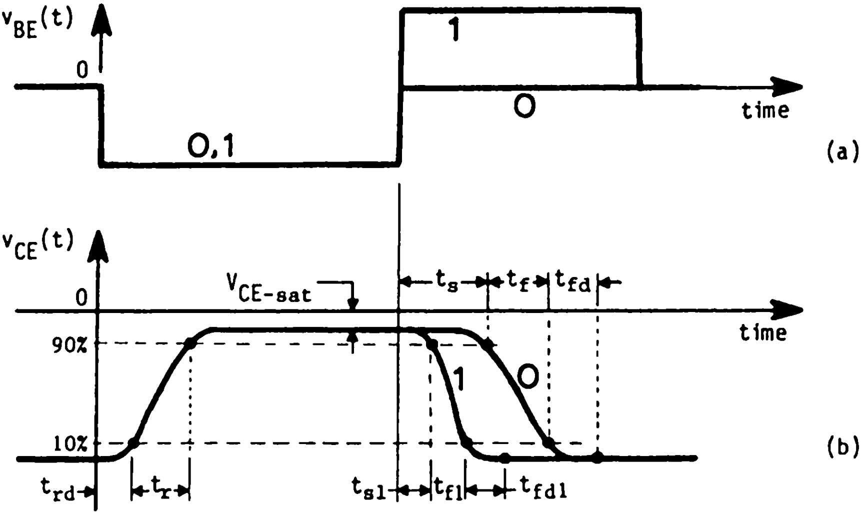


Fig.762.3 The waveforms of CE BJT inverter with overdrives. (a) Base input voltage. (b) Collector output voltage. Curve 0: zero turn-off overdrive. Curve 1: large turn-off overdrive.

(5) Discharging the Stored Base Charge in the Active Range

During this phase 5 the remaining base minority carriers (holes) injected by the forward biased emitter-base junction are extracted by the reverse biased collector-base junction. The governing charge control equations are the same as (753.1A) and (753.1B) for the turn-on active region when this remaining charge was injected. If a reverse base current, denoted by I_{B2}, is applied to speed up the extraction of the minority carriers (holes) stored in the base, the solution is then

$$i_C(t) = I_C - \beta_F(I_{B2} + I_C/\beta_F)[1 - \exp(-t/\tau_B)] \qquad (762.20)$$

$$= I_C\cdot\exp(-t/\tau_B). \qquad \{\text{for free turn-off with } I_{B2}=0\}. \quad (762.20A)$$

The 10% to 90% fall time is then

$$t_f = \tau_B\cdot\log_e\{[1+0.9(I_C/\beta_F I_{B2})]/[1+0.1(I_C/\beta_F I_{B2})]\} \qquad (762.21)$$

$$t_f = \tau_B\cdot\log_e 9 = 2.20\tau_B \qquad \{\text{for free turn-off with } I_{B2}=0\}. \quad (762.21A)$$

$$t_f \propto \tau_B\{0.8I_C/\beta_F I_{B2}\} = 0.8(I_C/I_{B2})\tau_B \qquad \{\text{if } I_{B2} >> I_C/\beta_F\}. \quad (762.21B)$$

Again, (762.21B) shows that overdrive speeds up the turn-off very significantly which is illustrated by curve 1 in Fig.762.3(b).

(6) Discharging the Space-Charge Layer Capacitances

The turn-off transient contains a final delay phase from discharging the space-charge-layer capacitance of the C-B and E-B junctions by majority carriers. The fall-delay time, t_{fd} or t_6, is similar to that of phase 1, about 0.15ns. This is small in 1-μm BJTs but significant in 0.1-μm BJTs. During this phase, the E-B junction voltage, $v_{EB'}(t)$, drops from the injection threshold ~0.6V to ~0V since the voltage drop due to the C-B junction leakage current passing through $R_B+r_{b'}$ is less than a few nV. Simultaneously, the C-B junction voltage, $v_{CB'}(t)$, drops from 0V to the static reverse value of $-5V+V_{EB'} \simeq -5V$.

Average Propagation Delay in CE BJT Inverter

BJT switches are specified by a figure of merit known as the average propagation delay. It is the average of the sum of the turn-on and turn-off delays for the output to reach 50% of its final value as illustrated in Fig.762.4. The 50% reference makes the numbers half as large as it should be and tends to give the impression of a faster transistor. (The 90% and 10% times, to be known as the pulse delay, t_{pd}, are preferred.) The 50% delay is defined by Fig.762.4 and the following equations. [See also (754.1)-(754.8) and Fig.754.1(d).]

$$t_p = \tfrac{1}{2}[(t_{rd} + \tfrac{1}{2}t_r) + (t_s + \tfrac{1}{2}t_f + t_{fd})] \qquad (762.22)$$

$$= \tfrac{1}{2}[(t_{rd} + t_{fd})]$$

$$+\tfrac{1}{2}\tau_B[0.8(I_C/\beta_F I_B) + \log_e(\beta_F I_B/I_C) + 1.1] \qquad (762.22A)$$

$$\simeq \tfrac{1}{2}\tau_B[0.08 + 2.3 + 1.1] = 1.8\tau_B. \qquad (762.22B)$$

Equation (762.22A) is for the free turn-off. It is derived using (762.7) and (762.7A) for t_r, (762.19A) for t_s, and (762.21A) for t_f.

To get (762.22B), it is assumed that the turn-on overdrive factor is 10, i.e., $\beta_F I_B/I_C \simeq \beta_F(V_{BB}/R_B)/(V_{CC}/R_C) = \beta_F(R_C/R_B) = 100(1k/10k) = 10$. It corresponds to the previous example of $R_B = 10k\Omega$ and $R_C = 1k\Omega$ if the internal series resistances are negligible. This example shows that the most important delays are the turn-off storage and fall times, contributing $2.3\tau_B/2$ and $1.1\tau_B/2$ respectively.

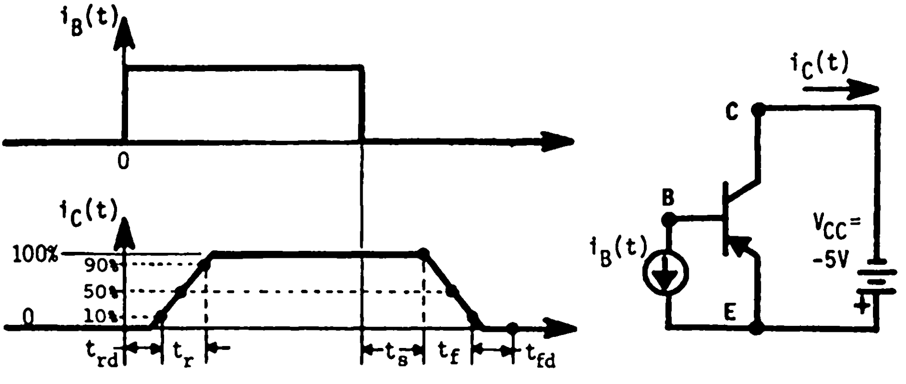


Fig.762.4 The waveforms for defining the average propagation delay of a CE p/n/p BJT inverter, roughly to scale for a -5V power supply.

763 Speeding Up the CE BJT Inverter

There are several ways to reduce the delays in a BJT inverter in the CE configuration. The speed-up methods are: (1) Large turn-on overdrive to reduce rise time and large turn-off overdrive to reduce both storage and fall times; (2) Input speed-up capacitor to reduce both rise, storage and fall times; (3) Schottky barrier diode clamp to prevent saturation; and (4) Emitter-coupled transistor drive to steer a less-than-saturation current source with limited voltage swing that prevents saturation. These are illustrated by circuit diagrams and waveforms. Detailed analyses are deferred to the next course on integrated circuits. The last two are known as the **nonsaturating** circuits while the first two, the **saturating** circuits.

(1) Large Turn-On and Turn-Off Overdrives

This was already analyzed and the waveforms were shown in Figs.762.3(a) and (b). The large turn-on would prolong the charge storage phase. A large turn-off overdrive requires a large reverse base input current which could be obtained from a CB first stage.

(2) Input Speed-Up Capacitor

The circuit and the sketched waveforms are shown in Figs.763.1(a)-(e) for p/n/p on the left and n/p/n on the right. The basic idea was already discussed with respect to Figs.753.2(a)-(c). The modification is the partial shunt of the input resistance by a capacitance C_B which provides a large initial base current pulse that will speed up the turn-on transient but still limits the steady-state stored base charge to the value without the capacitance. The turn-off transient is also sped up by the short-circuit action of the capacitance at the start of the turn-off transient.

(3) Schottky-Barrier Bethe Diode Clamp

The largest delay in the CE BJT inverter comes from the charge stored in the base during saturation because the collector-base junction is forward biased and injecting minority carriers into the base also. A even longer delay can be experienced if the collector quasi-neutral layer is thick and lightly doped or has high resistance. The underlap collector-base p/n junction diode will contribute more stored charge since it is also forward biased. One solution is to connect a metal/semiconductor rectifying diode (Schottky-barrier **Bethe diode**) in parallel with the collector-base junction. The reasons are: (i) the forward current in a m/s diode is completely carried by majority carriers and (ii) the threshold or cut-in voltage in a m/s diode is considerable lower than that of the collector-base p/n junction diode. Thus, the m/s diode clamps the collector-base forward voltage to such a low value that few minority carriers are injected into the base layer by the forward biased collector-base p/n junction.

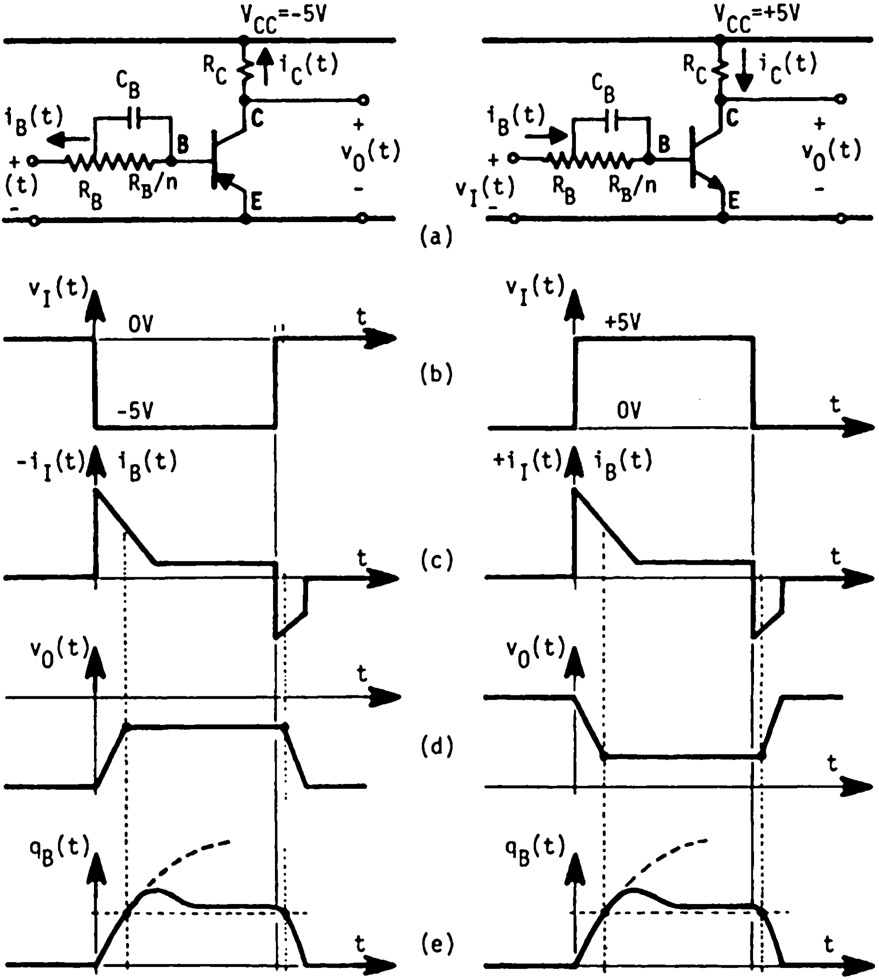


Fig.763.1 Speed up capacitance for CE p/n/p and n/p/n BJT inverters. (a) Circuit diagram and waveform of (b) input voltage, (c) input current, (d) output voltage, and (e) base charge.

A cross-sectional view, the circuit diagram and the circuit symbols of the SB Bethe m/s diode were given in section 565. The circuit diagram and the input and output voltage waveforms are shown in Fig.763.2. Two output waveforms are shown in Fig.763.2(c), one with and one without the m/s diode. There are two new features in the output waveform of the m/s-diode-clamped BJT: (1) the magnitude of the saturation voltage, $V_{CE\text{-sat}}$, is only a few tenth of volt larger than that without the m/s diode, so the output voltage swing is not reduced significantly, and (2) the storage delay is completely eliminated. The reasons are as follows. The analysis and data given in sections 562-563 showed that the m/s diode is much more conductive (1000 times or more) than a p/n diode at a given forward voltage. Thus, when the BJT is driven into saturation by a large base current, the forward voltage across the collector-base junction will be 'clamped' by the m/s diode to a low forward value and all the collector forward current will be conducted by the m/s diode. The low forward voltage gives a slightly larger magnitude of $V_{CE\text{-sat}}$ but it does not reduce the output voltage swing significantly. The m/s conducts by majority carriers, thus no minority carriers are injected into the base layer during the saturation phase and storage delay is completely eliminated.

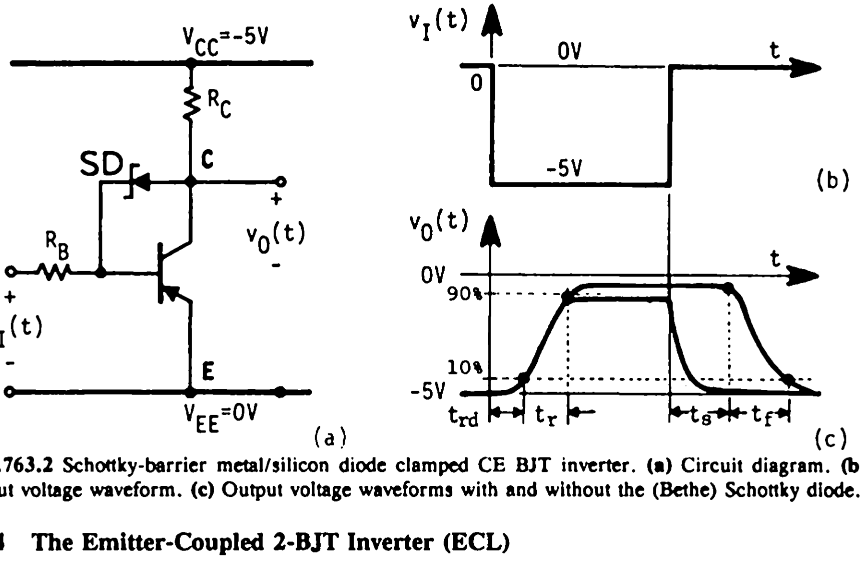


Fig.763.2 Schottky-barrier metal/silicon diode clamped CE BJT inverter. (a) Circuit diagram. (b) Input voltage waveform. (c) Output voltage waveforms with and without the (Bethe) Schottky diode.

764 The Emitter-Coupled 2-BJT Inverter (ECL)

The fourth speed-up method discussed at the beginning of the previous section is another **nonsaturating current switch**. It operates by steering a below-saturation collector current of the input BJT to a diode or a second transistor and by limiting the voltage amplitude so its excursion lies in the non-saturation or active region. The most successful is the emitter-coupled pair of two BJTs, known as the **emitter-coupled logic** (ECL) inverter. It was previously called the current mode logic (CML) in volume 6 of SEEC which gives a superb pedagogical description in section 2.2.2 on page 57. This shall be followed with updated notation and parameter values. The CML is an inverter which switches a constant current between two paths by the command of an input voltage. Figure 764.1(a) shows the basic circuit that uses a diode to steer the constant current I_{EE} between the diode and the transistor. The magnitude, $|I_{EE}|$, is selected to limit the collector current to less than saturation. Figure 764.1(b) shows the current path when the transistor is turned on and the diode is turned off. Figure 764.1(c) gives the voltage transfer characteristics. Figure 764.1(d) shows the practical ECL gate using a second transistor to provide the second current path and a emitter power supply and emitter resistance to give the constant current source.

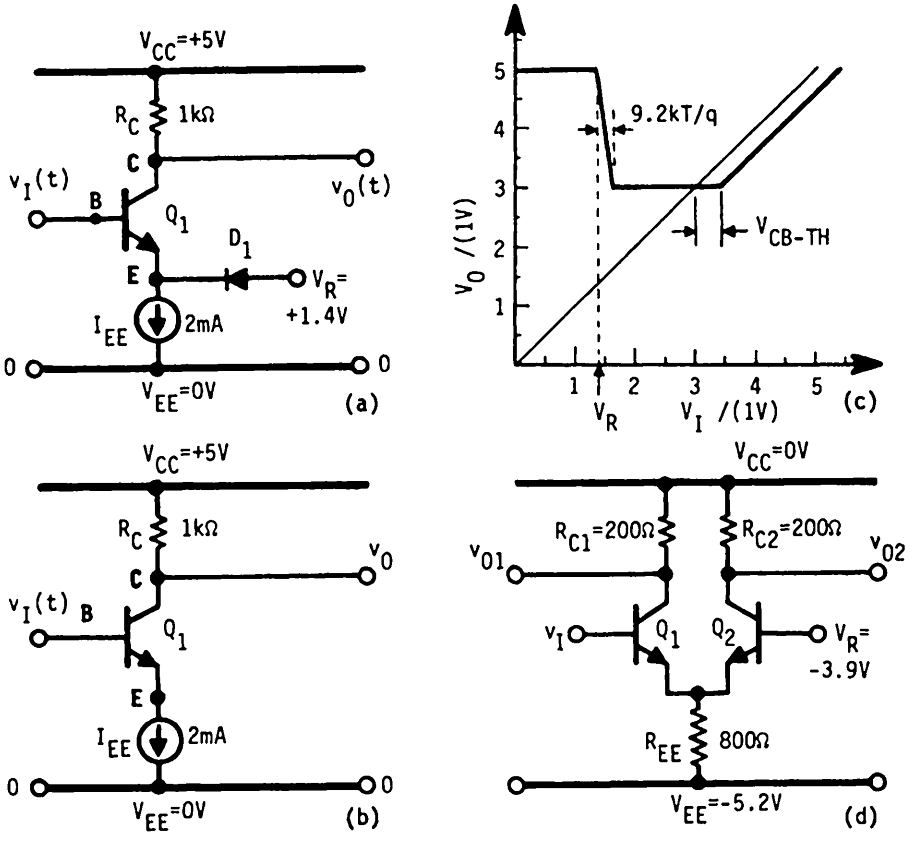

VCC=+5V
RC 1kΩ
C
vI(t)
B
Q1
vO(t)
D1
E
VR=+1.4V
IEE 2mA
0
0
VEE=0V
(a)
5
4
3
2
1
vO /(1V)
9.2kT/q
VCB-TH
1
2
3
4
5
VR
vI /(1V)
(c)
VCC=+5V
RC 1kΩ
vO
C
vI(t) B
Q1
E
IEE 2mA
0
0
VEE=0V
(b)
VCC=0V
RC1=200Ω
RC2=200Ω
vO1
vO2
vI
Q1
Q2
VR=-3.9V
REE 800Ω
VEE=-5.2V
(d)

Fig.764.1 The emitter-coupled logic (ECL) inverter. (a) The elementary circuit using a diode for the second current path. (b) Current path when the BJT is turned on. (c) Voltage transfer characteristics. (d) The emitter-coupled pair using a second transistor to provide the alternate current path. Note $V_R=-5.2+1.3=-3.9V=75\%$ of V_{EE} or 25% of V_{CC} if $V_{CC}=+5.2V$.

The circuit in Fig.764.1(a) operates as follows. Let us denote the n/p/n BJT by Q_1 and the diode by D_1. For ease of discussion, let $V_{CC}=+5V$, $V_R=+1.4V$ =reference, $R_C=1k\Omega$, $I_{EE}=2mA$, and $I_{ES}=0.2pA$ for both diode D_1 and the emitter-base junction of Q_1. These then give a forward junction voltage of

$$\begin{aligned} V_{EB} &= (kT/q)\log_e(2mA/0.2pA) = 23.0(kT/q) \\ &= 23.0 \times 0.02585 = 0.595V \simeq 0.6V. \end{aligned} \qquad (764.1)$$

This is higher than the injection threshold of all three diodes: $V_{EB\text{-}TH} \simeq V_{D1\text{-}TH} \simeq 0.5V$ and $V_{CB\text{-}TH} \simeq 0.4V$. The values are selected so that the electron-hole recombination current via recombination centers in the junction space-charge layers of the BJT and the diode can be neglected in this example.

When the base input voltage $v_I(t)$ is less than the reference voltage $V_R = +1.4V$ (Input is low or LO.), the BJT Q_1 is in the cut-off region because **its emitter-base junction is reverse biased**. Thus, the current from the constant current source, $I_{EE} = 2mA$, is steered to the diode D_1 and forward biases it to 0.6V. The collector current of Q_1 is then zero and the collector output voltage is at $V_{CC} = +5V$ (Output is high or HI). When $v_I(t)$ increases to $V^{BE}{}_{-Q1} + V_E + 9.2(kT/q) = V_{BE\text{-}Q1} + V_R - V_{BE\text{-}D1} + 9.2(kT/q) = V_R + 4.6(kT/q) = 1.4 + 0.2378V = 1.6378V$, the base-emitter junction of Q_1 is biased into strong forward conduction with 10^4 times higher conductance than D_1 [from $\exp[(q/kT)\cdot 9.2(kT/q)] = \exp(9.2) = 10^4$]. Thus, the 2mA from I_{EE} is now steered to Q_1 and the current through the diode D_1 is nearly zero. The current I_{EE} passes on to the collector and out through R_C as indicated by Fig.764.1(b) because the base current is low, $I_B = I_{EE}/(\beta_F + 1)$. This collector current causes a voltage drop of $I_{EE}R_C = 2mA\cdot 1k\Omega = 2V$ across the collector load resistance R_C and it decreases the output voltage from +5V to +3V as indicated in Fig.764.1(c). As the input voltage increases further and exceeds

$$v_I(t) = V_{CC} - I_{EE}R_C + V_{CB\text{-}TH} \qquad (764.2)$$

$$= 5 - 2mA\cdot 1k\Omega + 0.4 = +3.4V, \qquad (764.2A)$$

the output starts to rise again because the collector-base junction becomes heavily forward biased as indicated by $v_{BC}(t) = v_I(t) - v_O(t) = 3.4V - 3.0V = +0.4V$, and

$$v_O(t) = v_I(t) - v_{BC}(t) \qquad (764.3)$$

$$\simeq v_I(t) - V_{BC\text{-}TH}$$

$$= v_I(t) - 0.4V \qquad \{\text{Valid for } v_I(t) > 3.4V\} \qquad (764.3A)$$

shown in Fig.764.1(c). This would inject additional minority carriers into the base from the quasi-neutral collector. It is avoided by limiting the input to less than $V_{CC} - I_{EE}R_C = 3.0V$ in order to prevent the collector-base junction from going into forward bias and Q_1 into the saturation region. This non-saturation is the key to high-speed and makes ECL the fastest **non-saturating logic circuit**.

A practical implementation of the ECL inverter gate is shown in Fig.764.1(d). It employs a second transistor, Q_2, in place of the diode D_1; a negative emitter power supply $V_{EE} = -5V$ (commercial ECL gate arrays specify -5.2V); and an emitter resistor R_{EE} ($> R_C$) to provide both the emitter coupling between two transistors and the nearly constant current source, $I_{EE} \simeq V_{EE}/R_{EE}$. The reference voltage, V_R, is about 25% of the power supply voltage, $\simeq -1.3V$, obtained from a resistance or transistor voltage divider. The voltage swing is approximately $V_L = I_{EE}R_C \simeq [R_{C1}/(R_{EE} + R_{C1})]V_{EE}/5 = 1V$. The input voltage change to give a full output swing is about 200mV estimated from current steering above or

$$2(kT/q)\cdot\log_e(qV_L/kT) \simeq 2(kT/q)\cdot\log_e 40 \simeq 7.4(kT/q) = 191mV.$$

or

$$2(kT/q)\cdot\log_e(I_2/I_1) = 2(kT/q)\cdot\log_e 100 = 9.2(kT/q) = 238mV.$$

This is a rather narrow input voltage range to switch the state of the output. Thus, the ECL inverter is also known as a **threshold logic inverter**. Since ECL compares the input with a **reference voltage** to decide on its output state, it is also known as the digital **voltage comparator** and digital **difference amplifier**, although its gain is only slightly greater than unity. ECL circuit has two outputs: v_{O1} is the inverted output and v_{O2} is the non-inverted output so it can drive a balanced or push-pull output amplifier to further increase the load driving capability.

The intrinsic switching speed of the ECL inverter is limited by the rise and fall times of charging and discharging the base layer by minority carriers diffusing into and out of the base layer. However, the charging and discharging are limited to a small range of the active region, a voltage swing of 1V or about 20% of the 5V power supply. This limited swing shortens the rise and fall times proportionally, to about $(t_r+t_f)/5$. The rise and fall times, t_r and t_f, were given by (762.7) and (762.21) respectively. Thus, the ECL inverter is very fast, having an intrinsic switching speed faster than $0.8(I_C/I_{B1})t_B + 0.8(I_C/I_{B2})t_B \leq 10\cdot t_B$ if the external input and output resistances are the limiting factor. The 90% pulse delay is $K_d t_B = 2.73 t_B$. If these external resistances are small, then the pulse delays in ECL switching circuits are mainly from charging and discharging the space-charge layer capacitances and the parasitic and load capacitances, C_L, through the internal collector and base resistances. This has a time constant of C_L/g_m, just like the MOST circuits described in section 663, but g_m of a BJT is 10-to-100 times larger than MOST, making the BJT correspondingly faster. Operation characteristics of various improved ECL inverters are described in digital circuit textbooks. [Such as David A. Hodges and Horace G. Jackson, Analysis and Design of Digital Integrated Circuits, 2nd ed. McGraw-Hill Book Co., New York, 1988.]

765 The CB-CE Transistor-Transistor Coupled 2-BJT Inverter (TTL)

The second two-transistor BJT inverter is the transistor-transistor inverter known as the **transistor-transistor logic** (TTL). It uses a CB input transistor to drive a CE output transistor. It has been historically the most popular BJT inverter after a family of TTL logic gates was introduced by Texas Instruments in 1965. One factor for its selection to build the logic family was the simple device geometry of multiple input gates for monolithic integrated circuits. It was known as the **Phoenix Gate** suggested by the multiple input leads shown in the circuit diagram given in Fig.765.1(a) for a 4-input TTL NAND gate. A second factor was the ease of fabrication shown by the cross-sectional view in Fig.765.1(b). The multiple input emitters (four illustrated) are all obtained in one diffusion step on a common base layer. The input and output resistors are obtained from an extension of the diffused p-base and n-collector. The 2-BJT circuit in Fig.765.1(a) also shows that both an inverted output, NAND, and a non-inverted output, AND, are available. Figure 765.1(c) shows that these outputs are used to drive a **totem-pole** output circuit with Q_3 sitting on top of D_1 and Q_4 like a totem-pole. It is also known as the **push-pull (push-down** or pull-down and **pull-up)** output circuit. Q_3 and $Q4$

greatly reduce the output resistance of both the high and low output states and improve the speed when driving a large capacitance load. The output resistance is reduced roughly from $R_C = 1k$ of Q_2 in Fig.765.1(a) to the following values in Fig.765.1(c): (1) $R_{CE\text{-}sat4} = 15\Omega$ of Q_4 at output low, and (2) $(kT/qI_{D1}) + R_C/\beta_{F3} \approx (26mV/1mA) + (1600/40) = 66\Omega$ at output high. Q_4 actively pulls (pushes) down the output voltage while Q_3 actively pulls up the output voltage.

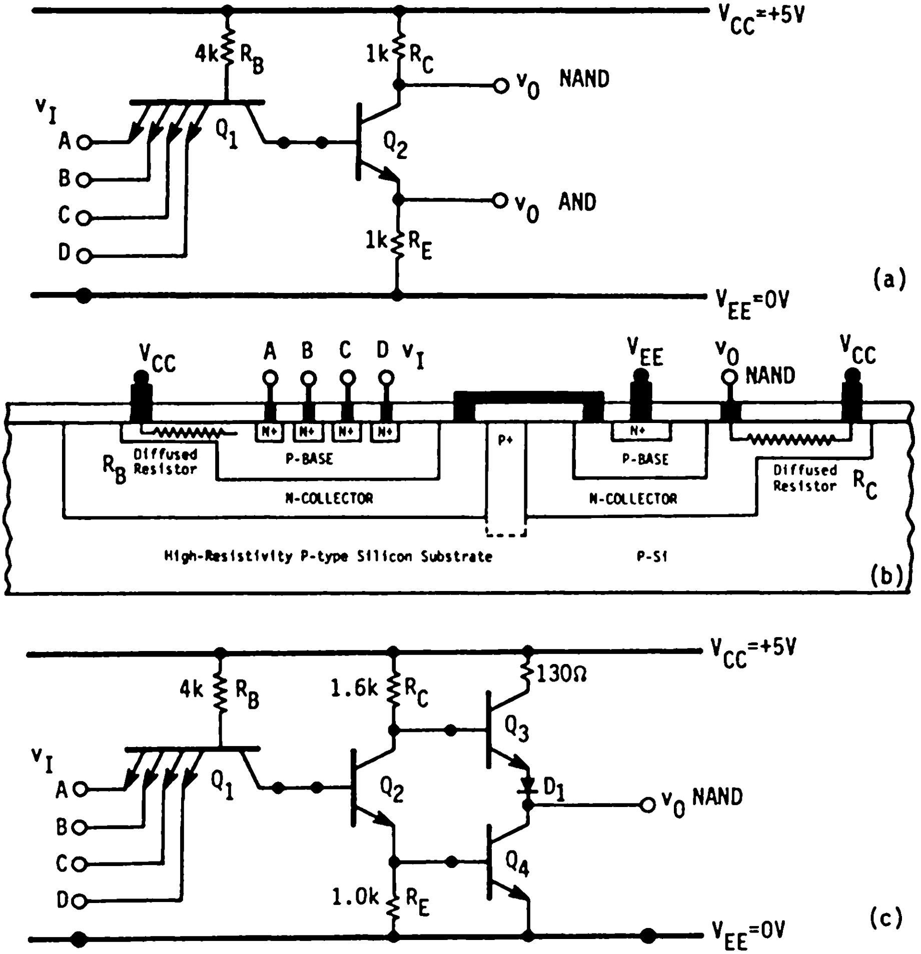

V_CC=+5V
4k
R_B
1k
R_C
v_O NAND
v_I
A
B
C
D
Q_1
Q_2
v_O AND
1k
R_E
V_EE=0V
(a)
V_CC
A B C D v_I
V_EE
v_O NAND
V_CC
N+
R_B Diffused Resistor
P-BASE
N-COLLECTOR
P+
P-BASE
N-COLLECTOR
Diffused Resistor R_C
High-Resistivity P-type Silicon Substrate
P-Si
(b)
V_CC=+5V
4k
R_B
1.6k
R_C
130Ω
Q_3
v_I
A
B
C
D
Q_1
Q_2
D_1
v_O NAND
Q_4
1.0k
R_E
V_EE=0V
(c)

Fig.765.1 The transistor-transistor logic (TTL) circuit. (a) The basic circuit diagram. (b) The cross-sectional view. (c) With totem-pole output circuit. For visual ease, the thickness of the diffused resistors are shallower than the base or collector. Later TTL uses epitaxial or ion implant layer of uniform thickness for collector and a smaller uniform thickness for base.

As indicated in section 754, when the BJT is driven by a current or open-circuit source during turn-off, the signal delay in the CB configuration (t_B) is much smaller than that in the CE configuration ($\tau_B = \beta_F t_B$). Thus, the intrinsic speed of the 2-transistor TTL inverter is limited by the carrier recombination lifetime, τ_B, in the base of the output transistor Q_2 since it is in the CE configuration. The signal delay through the input BJT Q_1 is much smaller since it is in the CB configuration. However, the turn-off delay in Q_2 is reduced considerably when the emitter input voltage of Q_1 drops to zero because the emitter current of Q_1 must now flow out of the base of Q_2, pulling out the stored charge in the base of Q_2. This is the turn-off overdrive in the CE inverter discussed in section 763 (1), and described by $q_S(t)$ in (762.18), t_s in (762.19), $i_C(t)$ in (762.20) and t_f in (762.21).

The methods described in 763 can also be used to speed up the turn-off transient. The most popular approach is to use a Schottky-barrier diode to prevent the output transistor Q_2 from being driven into saturation but its high leakage doubles the power dissipation which was then lowered by 5X increase of the resistances via finer lithography. The chronology in Table 765.1 shows the speed increase and power reduction from newer technology.

Using the 12-micron lithography, the 74xx (xx=2 or 3 digit part number) logic family in the first column of Table 765.1 was introduced around 1965 with 10ns average propagation delay per gate and 10mW power dissipation, giving a delay-power product of 10ns•10mW=100pJ. The 74Sxx TTL logic family using Schottky-Bethe diode clamp and the same 12-micron lithography was introduced five years later in 1970 which increased the speed to 3ns but also increased the power dissipation by 2x to 20mW, giving an improved delay-power product of 3ns•20mW=60pJ. It took another five years until 1975 to put the new 6-micron lithography into production to give the 74LSxx low power Schottky TTL gates with 2mW dissipation and 10ns speed or power-delay product of 20pJ. In 1980, the 3-micron technology was put into production to give the fast (74Fxx), advanced Schottky (74ASxx), and advanced low-power Schottky (74ALSxx) families with power-delay products of 4x2.5=10pJ, 1.5x10=15pJ, and 2x4=8pJ respectively.

The TTL families will be completely replaced eventually by the much lower power and faster CMOS gates with BiCMOS drivers (BiCMOS, see the next section), and the much faster ECL gates, both listed in Table 765.2. The 74HCxx and 74ACxx CMOS and advanced CMOS technologies introduced in 1985-1986 gave power-delay products of 2.5μWx10ns=25fJ and 2.5μWx5ns=12fJ without BiCMOS drivers. The 100k ECL gates introduced in 1980 had a power-delay product of 40mWx750ps=30pJ. The latest ECL and NTL (none threshold logic) are approaching 1mW, 10ps, and 10fJ. This table shows that Si CMOS with BiCMOS driver will dominate medium to high speed (100ps) and low power (1μW/gate) applications with power-delay products less than 1pJ, while Si bipolar ECL, having a power-delay product greater than 1pJ, will only be used only in very high speed applications (< 10ps) due to higher power dissipations (> 1mW).

Table 765.1
Specification of TTL Gates
(T_A=25°C)

PART NUMBER		74	74S	74LS	74F	74AS	74ALS
INTRODUCTION YEAR		1965	1970	1975	1980	1980	1980
minV_{OH}/MaxV_{OL}	(V/V)	2.4/.4	2.7/.5	2.7/.5	2.7/.5	2.7/.5	2.7/.5
minV_{IH}/MaxV_{IL}	(V/V)	2.0/.8	2.0/.8	2.0/.8	2.0/.8	2.0/.8	2.0/.8
minI_{OH}/minI_{OL}	(-mA/mA)	.4/16	1/20	.4/8	1/20	2/20	.4/4
MaxI_{IH}/MaxI_{IL}	(µA/-mA)	40/1.6	50/2	20/.4	20/.6	200/.2	20/.2
Logic Swing	(V)	3.3	3.3	3.3	3.3	3.3	3.3
Power Supply	(V)	+5.0	+5.0	+5.0	+5.0	+5.0	+5.0
Power Dissipa	(mW/gate)	10	20	2	4	20	2
Propag. Delay	(ns)	10	3	10	2.5	1.5	4
Fan-out		10	10	10	10	10	10
Si Technology		NPN Standard	NPN Schottky	NPN LoP.Sc	NPN Fast	NPN Ad.Sc	NPN+PNP Ad.LoP.Sc
Line Width	(µm)	12	12	6	3	3	3
Oxide/Base	(nm)						

Table 765.2
Specification of CMOS and ECL Gates
(T_A=25°C)

PART NUMBER		74HC	74AC	10K	100K	EXPERIMENTAL	
INTRODUCTION YEAR		1985	1986	1971	1980	1987	1991
minV_{OH}/MaxV_{OL}	(V/V)	4.4/.1	4.4/.1	-0.9/-1.7			(2Mgate)
minV_{IH}/MaxV_{IL}	(V/V)	3.1/.9	3.1/1.3	-1.2/-1.4			
minI_{OH}/minI_{OL}	(-mA/mA)	4/4	24/24	22/22			
MaxI_{IH}/MaxI_{IL}	(µA/-mA)	.1/.1	.1/.1	12µ/12µ			
Logic Swing	(V)	4.7	4.7	0.8	0.8	0.45	3.3
Power Supply	(V)	+5.0	+5.0	-5.2	-5.2	3	3.3
Power Dissipa	(*/gate)	2.5µW	2.5µW	24mW	40mW	10/2mW	<1µW
Propag. Delay	(ps)	10k	5k	2k	<750	30/20	230ps
Fan-out		10	10	10	90	1	7
Si Technology		CMOS	CMOS	ECL	ECL	ECL/NTL	BiCMOS
Isolation				p/n	SiO_2	Self-Align	
Line Width	(µm)	3	2			0.35x5	0.5
Oxide or Base	(nm)	60	40			70	<10

**SOURCES:*

The TTL Data Book, Texas Instruments.
Fast and LS TTL Data, Rev.4 (01/89), **MECL System Design Handbook**, Motorola
CMOS Data Book, Fairchild now National Semiconductor Corp.
ISSCC Technical Digest up to and including 1991.

766 The BIPOLAR-MOS Inverters (BiMOS, BiCMOS, CBiCMOS)

It was demonstrated in chapter 6 that MOSTs and all FETs have a low transconductance or g_m compared with BJTs described in this chapter. The difference is essentially equal to the ratio of the thermal voltage to the operating voltage at the same operating current or current density, $I_D(\text{drain}) = I_C(\text{collector})$,

$$g_m(\mathrm{MOS})/g_m(\mathrm{BJT}) \approx (I_D/V_{DD})/[I_C/(kT/q)] = (kT/qV_{DD})$$

$$\approx 0.025/5.0 = 0.005 = 1/200. \qquad (766.1)$$

This comes about as follows. The load current or the current carrier density in BJT varies very nonlinearly and extremely rapidly, i.e., exponentially with the controlling voltage following the Boltzmann factor, exp(qV/kT), because BJT operates by modulating the load current or carrier concentration via controlling the potential barrier height of a p/n junction and hence the carrier density (**Shockley's minority carrier density modulation and injection principle**). The load current or carrier density in MOSFET and all FETs varies with the control voltage rather slowly, as V^n where $n=1$ to 2, because FET operates as a variable resistance or conductance modulated by a capacitively coupled electric field perpendicular to the length direction of the conductor (**Lilienfeld's conductivity modulation principle**). For this fundamental reason which gives the large transconductance difference, the **extrinsic** or **circuit speed** of a BJT is much faster than a MOSFET even when they have the **same intrinsic speed** because the BJT has a much larger transconductance to charge and discharge or **drive** the capacitance of the circuit load. In fact, their intrinsic speeds are comparable at a given submicron dimension because of the smaller input capacitance of the MOS gate oxide of the intrinsic FETs compared with the large input capacitance of the forward-biased emitter-base junction of the BJT. The comparable intrinsic speed which is proportional to the intrinsic cutoff frequency, ω(device-intrinsic), can be illustrated by the following comparison:

$$\omega(\mathrm{MOST\text{-}intrinsic}) = g_m(\mathrm{MOS})/(C_{gs}+C_{ds}) \qquad (766.2A)$$

$$\approx \omega(\mathrm{BJT\text{-}intrinsic}) = g_m(\mathrm{BJT})/(C_{eb}+C^{c}) \qquad (766.2B)$$

despite

$$g_m(\mathrm{MOS}) \ll g_m(\mathrm{BJT}) \qquad (766.3A)$$

because

$$(C_{gs}+C_{ds})(\mathrm{MOS}) \ll (C_{eb}+C_{ds})(\mathrm{BJT}). \qquad (766.3B)$$

The slow circuit speed of an intrinsically very fast MOSFET or any FET is caused by its low g_m to charge and discharge large load capacitances from interconnect wiring, C_L. This makes ω(MOS-circuit) << ω(BJT-circuit) since

$$\omega(\mathrm{MOS\text{-}circuit}) = g_m(\mathrm{MOS})/(C_{gs}+C_d+C_L) \approx g_m(\mathrm{MOS})/C_L \qquad (766.4A)$$

$$\ll \omega(\mathrm{BJT\text{-}circuit}) = g_m(\mathrm{BJT})/(C_{eb}+C_c+C_L) \approx g_m(\mathrm{BJT})/C_L. \qquad (766.4B)$$

This is the major speed limit of high density MOS integrated circuits whose on-chip (aluminum) metal-on-oxide transmission lines for interconnecting transistors on the chip are long and give high capacitances. Lower power dissipation and easier manufacturing (hence higher yield and lower cost) are the main reasons of preference for MOS over bipolar in high density integrated circuit chips for medium speed applications.

Several approaches have been developed to reduce and finally overcome the low g_m limitation of MOST in order to increase the integrated circuit speed. The technology evolution to get larger transconductance shows roughly the following historical sequence each with the indicated drawbacks. (1) Increase the aspect ratio (W/L) of the line driving MOST (N-channel for higher electron mobility than hole) which increases the power dissipation and uses more area or reduces transistor density (transistor or function per chip). (2) Use CMOS line driver with large aspect ratio to give high g_m at negligible additional standby power dissipation but which still uses more area or reduces transistor density. (3) Use BJT line driver, known as bipolar-MOS or **BiMOS**, which increases power dissipation but uses a much smaller area. (4) Use two BJTs as a push-pull line driver for a CMOS, known as **BiCMOS**, which uses slightly larger area than (3) but reduces power dissipation to a negligible level. This 4-transistor (4-T) inverter has begun to appear in dense high-speed Si integrated circuit products in 1990. (5) Use a complementary BJT push-pull line driver pair for a CMOS, known as **CBiCMOS**, which further increases the speed to drive load capacitances. Use of CBiCMOS driver in very high speed and dense logic circuits has just begun in engineering reports and articles during 1989-1990. (See Proceedings IEDM and ISSCC.) Production of CMOS-CBJT chips is delayed by the slower speed of p/n/p BJT than n/p/n BJT which is being overcome by vigorous research efforts. The fastest p/n/p at the end of 1989 was reported by IBM Research Center at the December-1989 IEDM which gave the following performance figures of a self-aligned poly-emitter p/n/p Si BJT with $0.5\text{x}4.0\mu m^2$ emitter and 80nm thick base: $f_{Max}=27GHz$; $t_{pd}=18.4ps$ at $1.14mA/\mu m^2$ as determined by the propagation delay in an active pull-down ECL circuit with FI=FO=1 and 0.5V swing; and a NTL (non-threshold logic) gate delay of 36ps at $V_L=300mV$ and $1mA/\mu m^2$. It dissipated 50% less power than conventional ECL. Updates in 1991 gave $f_{Max}\rightarrow 50GHz$ and $t_{pd}\rightarrow <10ps$. The interconnect delay is now the final speed limit.

In the following paragraphs, we shall describe pedagogically the technology developments listed in (3), (4) and (5). Several alternate routes to reach (5) will be given. This pedagogical rather than historical-chronological presentation helps to understand the physics and systematize the circuit design and physical implementation methodologies.

The idea of using an output BJT to increase the transconductance or the load-driving capability of a MOST (or any FET and other electron devices that have a low transconductance) was well-known to circuit designers and those familiar with the semiconductor-device art. New circuits and experimental demonstration were reported in the literature at least a decade before silicon transistor technology was sufficiently mature to enable Hung Chang Lin (currently Professor of Electrical Engineering at the University of Maryland) to invent and implement the BiCMOS circuit monolithically on Si in late 1969 at the Westinghouse laboratory. The combination of an input MOST with an output BJT has resulted in the achievement of the most desirable characteristics of an amplifying or switching electron device: almost open-circuit input impedance (the MOS capacitance of the gate oxide) and almost short-circuit output impedance, given by $(r_{ch}+r_{b'})/\beta_F$.

To analyze the basic BiMOS circuits systematically, a pedagogical approach is used by considering all the combinations of transistor, circuit and bias which gives 22 workable 2-T and 4-T basic BiMOS circuits that give either inverted or not inverted output. The variations are: (i) the four (4) transistors, n-channel and p-channel MOSTs, n/p/n and p/n/p BJTs; (ii) the two (2) circuit connections between the MOST and the BJT in a pair: MOST connected in parallel with (or shunting) the emitter-base junction of the BJT, and MOST connected in series with the base terminal and the power supply (which has the topology of paralleling or shunting the collector-base junction but it is for controlling the base current and not shunting or diverting the collector current); and (iii) the two (2) MOST bias conditions: the reverse-biased (normal) and forward-biased (historical first BiMOS) MOST drain/substrate junction (or source/substrate junction). The enhancement and depletion modes of the MOST is not distinguished to simplify the following analysis by assuming that a proper d.c. bias is applied to the gate to give the necessary enhancement mode or depletion mode condition. The connection of the substrate terminal is crucial and is explicitly indicated in the circuit and monolithic physical realizations (cross-sectional views).

The twenty-two circuits are: (A) twelve 2-T BiMOS circuits consisting of (A1) eight BiMOS circuits with the MOST shunting the EB junction, and (A2) four 2-T BiMOS circuits with the MOST connected in series with the base terminal; and (B) ten 4-T BiCMOS circuits consisting of: (B-1) two CBiCMOS's, one NBiCMOS, and one PBiCMOS which uses the MOST to shunt the EB junction and divert or steer the emitter or base current, and (B-2) two each of the NBiCMOS, PBiCMOS, and CBiCMOS circuits which use the MOST in series with the base terminal to source the base current, which is also topologically in parallel with the CB junction in some of the BiCMOS circuits but is not intended to shunt or divert the collector current such as the ECL inverter.

The basic inverter circuit further proliferates when the biasing and level setting resistances between the BJT and MOST are included in the systematic evolution. Illustrations showing the need of these resistances will be given to describe the evolution of the circuits. Additional level setting, speedup, pullup, and pulldown resistors, diodes, and MOS transistors will be included in some circuits to illustrate their functions such as to achieve full-swing or rail-to-rail output (like the

CMOS) so the output amplitude is equal to the sum of the d.c. power supplies. The BiMOS circuits with EB-shunt using forward and reverse drain biases, and with Base-series using reverse drain bias are described in the following two subsections and their sub-subsections.

Shunt-BiMOS: Emitter-Base Junction Shunted by MOST

This circuit configuration and its monolithic realizations will be known as the **EB-shunt BiMOS** (the emitter-base junction of a BJT is shunted by a MOST). There are two families: the drain junction of the MOST is (i) forward biased (the non-conventional MOST operation mode), and (ii) reverse biased (the normal MOST operation mode). The high-gain EB-shunt BiMOS circuits evolved from a fundamental study undertaken by this author during 1960-1961 concerning the effects of surface recombination and surface channel on p/n junction diode and BJT characteristics [766.1]. In that study, two structures were used: **(a) one gated p/n junction diode** consisting of a p/n junction with a MOS gate over the surface intercept of the p/n junction [766.1]; and **(b) two gated p/n junction diodes** [766.2] that share a common gate, i.e. the modern MOST, in which one of the gated diode is the emitter-base junction of a BJT while the second diode under the gate is used to terminate the surface channel [which is floating in structure (a)] in order to increasing the channel current and gain for applications [766.3]-[766.4]. The latter structure (b) was also extended to the two emitter-base junction of p/n/p/n silicon controlled rectifiers (SCR) giving it essentially open-circuit input impedance for power control applications described briefly in section 739. An evolution history of the EB-shunt BiMOS was described in a recent review [766.5].

The two structures, (a) and (b), must be distinguished. Recent (1989-1990) literature from IBM researchers on new CBiCMOS circuits have termed the MOST in (b) a gated diode [which is the name for (a) and not (b)]. This is a misnomer which will be avoided in this book because the second p/n diode in (b) is crucial in the CBiCMOS and other applications since the second p/n diode is required to give high gain, while structure (a) gives not only low gain but the gain depends on unreliable, irreproducible and noisy surface recombination which must be avoided.

[766.1] C.T.Sah, "Effects of surface recombination and channel on p-n junction and transistor characteristics," IRE Trans.Electron Devices, ED-9(1), pp.94-108, Jan.1962.

[766.2] C.T.Sah, "Surface-potential controlled semiconductor devices," U.S.Patent 3204160, application filed Apr.12,1961, granted Aug.31,1965; U.S.Patent 3243699, application filed June 11,1962, granted Mar.29,1966. See also "Transistors," in 1963 McGraw-Hill Yearbook of Science and Technology, New York, NY: McGraw-Hill, pp.560-562.

[766.3] C.T.Sah "A new semiconductor tetrode - the surface-potential controlled transistor," Proc. IRE, 49(11), 1623-1634, Nov.1961.

[766.4] H.Z.Bogert, C.T.Sah and D.A.Tremere, "Applications of the surface-potential controlled transistor tetrodes," Proc. ISSCC, 34-35, Feb.1962.

[766.5] See section IV.H. on p.1294 of C.T.Sah, "Evolution of the MOS transistor - from conception to VLSI," Proc.IEEE, 76(10), 1280-1326, Oct.1988.

The circuit and monolithic implementation of the nMOST-n/p/n BiMOS are shown in Figs.766.1(a)-(d). This circuit has not been used alone in integrated circuit chips due to large standby power from the shunting current. However, it is ingenuously used to attain full-swing of the output voltage in BiCMOS and CBiCMOS (Complementary Bipolar and CMOS) monolithic circuits. Furthermore, a systematic pedagogical evolution of this circuit also leads to the 1969 zero-standby-power BiCMOS of H.C.Lin. Thus, the physical principle of operation and the topology of the electrical circuit and monolithic realization on Si of this EB-shunt BiMOS circuit and its derivatives will be described in detail qualitatively in the following paragraphs using these figures.

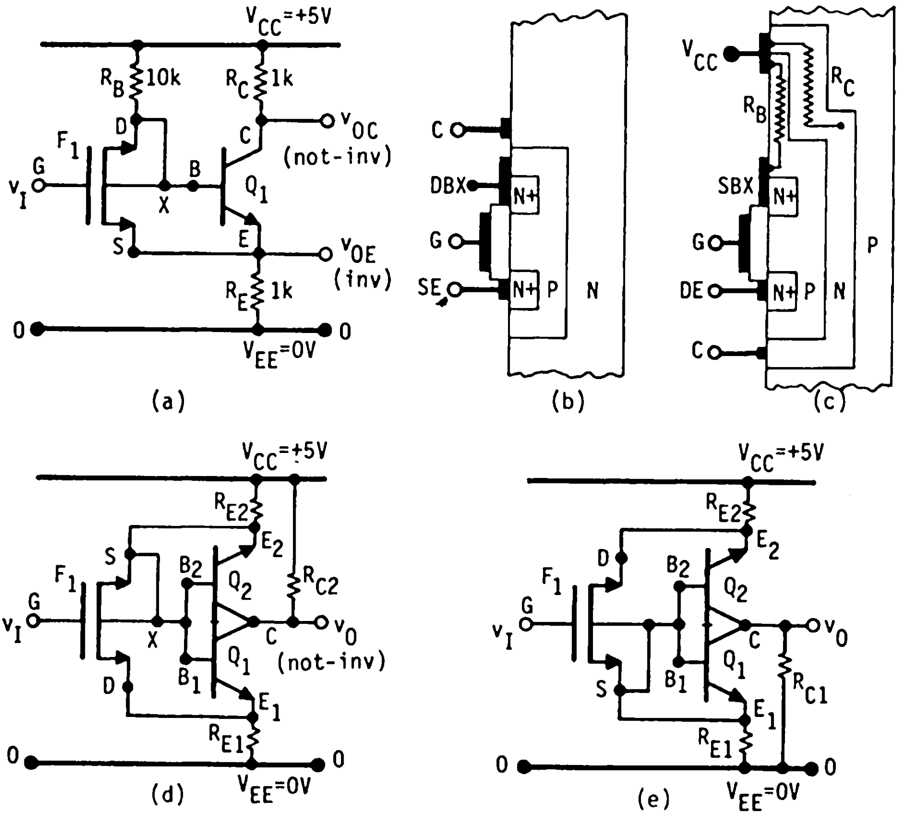

V_CC=+5V
R_B 10k
R_C 1k
D
C
V_OC
(not-inv)
F_1
G
V_I
B
X
Q_1
S
E
V_OE
(inv)
R_E 1k
0
0
V_EE=0V
(a)
C
DBX
G
SE
N+
N+
P
N
(b)
V_CC
R_C
R_B
SBX
N+
G
P
DE
N+
P
N
C
(c)
V_CC=+5V
R_E2
E_2
S
B_2
F_1
Q_2
R_C2
G
V_I
X
C
V_O
Q_1
(not-inv)
D
B_1
E_1
R_E1
0
0
(d)
V_EE=0V
V_CC=+5V
R_E2
E_2
D
B_2
F_1
Q_2
G
V_I
C
V_O
Q_1
S
B_1
R_C1
E_1
R_E1
0
0
(e)
V_EE=0V

Fig.766.1 Evolution of the MOS-bipolar integration which began in 1961. (a) Basic circuit. (b) Cross-sectional view of BiMOST alone. (c) Cross-sectional view with biasing diffused-resistors which are illustrated with a change of layer thickness while later epitaxy and ion implant technologies gave uniform thickness. (d) Circuit with both BJTs shown. (e) The base-source short-circuits replaced by MOSTs.

Figure 766.1(a) shows the basic circuit. The monolithic realization is shown in figure (b) which is an nMOST in a p-type base-well with two n+ regions as the source and the drain whose designation can be interchanged to simplify physical understanding and circuit-device analysis. This was the simplest experimental structure to build in 1961 because the standard four-mask set of the planar silicon BJT technology at that time could be used. Without the short-circuited upper n+/p junction in figure (b), it was also the most relevant structure for studying surface effects on an isolated p/n junction or on a BJT by varying the surface potential with the voltage applied to the MOS gated n+/p junction.

The lower n+ island in figure (b) also serves as the n+ emitter of the BJT, so it can be called the source of the MOST because it is the 'source' of the minority carriers (electrons) of the n/p/n BJT as well as the source of minority carriers (electrons) of the n-channel MOST. (In a later analysis, the source-drain label is interchanged to simplify circuit-device analysis.) Using this convention (source=emitter), the upper n+ island is then the drain and it is short-circuited to the p-type base or 'well-substrate' labeled X, i.e. $V_D=V_X=V_B$. Since the n+/p emitter-base junction is forward biased by $V_{CC}=+5V$ through the base resistance $R_B=10k\Omega$, the source-substrate n+/p junction (lower n+ island) is also forward biased. The forward bias is around 0.8V. Assuming a $r_{b'}$ drop of 0.2V, then the base/emitter and substrate/source forward bias voltage is about $V_B=V_X=V_D = 0.8+0.2 = 1.0V$. Then, the current flowing through the base biasing resistance $R_B=10k\Omega$ is $I_B=(5-0.8-0.2)V/10k\Omega=0.4mA$, giving $r_{b'}=0.2V/0.4mA=500\Omega$.

The circuit operation will now be described for the following two circuit configurations: (i) collector output ($R_C\neq 0$ and $R_E=0$), and (ii) emitter output ($R_C=0$ and $R_E\neq 0$).

(i) Collector Follower

Suppose that the nMOST is an enhancement mode device with $V_T=0V$. Then at $V_G=0$(INPUT LOW), we have $V_G - V_S - V_X = 0 - 1 = -1V$ so that the nMOST is off or non-conducting. Thus, all the 0.4mA R_B current is flowing into the base lead of the n/p/n BJT, turning it on. Suppose that $R_C=100\Omega$ for a high-capacity BJT driver, then $I_{Cmax} = (V_{CC} - V_{CEsat})/R_C = (5 - 0.1)/0.1k = 49mA$. So if $\beta_F < 49mA/0.4mA = 122$, then the BJT is not in saturation. Generally, $V_{C\text{-}out}$ = (the larger of $V_{CC} - \beta_F I_B R_C$ and $V_{CE\text{-}sat}$) $\approx$ 0.1V(OUTPUT LOW) and the output impedance is R_{sat}.

However, when $V_G=+5V$(INPUT HIGH), we have $V_G - V_S - V_X = 5\text{-}0\text{-}1 = 4V$ so the nMOST is turned on. Assume that the on-conductance of the nMOST at $V_G - V_X - V_S = +4V$ is much larger than the d.c. conductance of the BJT which is approximately $0.4mA/0.8V=0.5mS$. Then, the nMOST will divert the entire base biasing current ($I_B=0.4mA$) from the BJT's base lead to the nMOST channel.

Thus, the BJT is now turned off so that $V_{C\text{-out}} = +5V$(OUTPUT HIGH) with an output impedance (resistance) equal to R_C or 100Ω. In practice, the EB-shunt nMOST conductance may not be much larger than 0.5mS, then only a fraction of I_B is diverted from the BJT to the nMOST. In order to get a large diversion and hence high gain, one must use a lower base drive or larger R_B to lower the I_B.

Thus, the collector output is not inverted; it is a **collector follower**. This is expected since both the drain output of the nMOST and the collector output of the nBJT are inverted outputs, thus, two series inversions give noninversion.

(ii) Emitter Inverter

A similar consideration for case (ii) shows that the emitter output is inverted with OUTPUT = LOW = 0V when INPUT = HIGH = +5V and OUTPUT = HIGH = $V_{CC} - V_{CE\text{-sat}} \simeq 5.0 - 0.1 = 4.9V$ when INPUT = LOW = +0V.

The transconductance is approximately that of the nMOST multiplied by β_F or approximately $G_m \simeq (I_B/V_B)\beta_F = (0.4mA/1V)\text{x}122 = 50mS$ if the BJT is not driven into saturation. The intrinsic delay is mainly from t_B since the B-E junction is essentially short-circuited during turn-off provided $r_{b'} + R_E$ is not large. If it is large, then base recombination delay will prolong the turn-off transient.

The biasing resistance R_B and the load resistance R_C and R_E can be implemented monolithically as indicated in Fig.766.1(c). The full circuit diagram is given in figure (d) which shows the two BJTs of the physical structure of figure (b).

The foregoing descriptions suggest the following three extensions.

(1) There are four possible 2-T configurations which combine the forward-biased source/substrate junctions from the nMOST and pMOST with the two BJTs, the nBJT (n/p/n or electron current dominant) and pBJT (p/n/p).

(2) The notation choice of designating the forward-biased emitter n+region as the source with the drain n+region short-circuited to the p-type substrate-base well is not unique. The opposite choice is that of interchanging the source-drain designation, i.e., the emitter n+region is the forward-biased drain while the n+source is short-circuited to the p-type substrate and the base. This reversal of the source and drain designations exemplifies two important points: (a) the threshold voltage is now measured with respect to the source which is short-circuited to the substrate-well so that there is clearly no substrate bias, $V_X - V_S = V_{XS} = 0$; and (b) the n+drain is forward biased showing the distinctly different but rarely used forward MOST I_D-V_D curves in the third quadrant of Fig.643.1(a). These forward I_D-V_D curves are concave upwards and hence give a larger channel current $I_{CHF} = -I_D > 0$ and a larger total transconductance, $G_M = |I_{CH}/(V_G - V_{GT})|$, than the reverse or normal characteristics in the first

quadrant. This difference can be demonstrated mathematically by the simple 1-d MOST d.c. equation, (643.3), which was derived by using $V_X = V_S = \text{reference} = 0$ in the general result, (643.2). For normal (reverse drain bias) we have $V_{DS} = V_D - V_S = V_D > 0$ and $I_{CHR} = I_{DS} > 0$ and for this rarely used forward drain bias operation, we have $V_{SD} = (V_S - V_D) = -V_D > 0$ and $I_{CHF} = I_{SD} = -I_D > 0$. The equations from (643.3) are then

$$I_{CHR} = (Z/L)\mu_n C_o[(V_{GS}-V_{GST})V_{DS} - \tfrac{1}{2}V_{DS}^2] \quad \textbf{Reverse Drain Bias} \tag{766.5A}$$

$$I_{CHF} = (Z/L)\mu_n C_o[(V_{GS}-V_{GST})V_{SD} + \tfrac{1}{2}V_{DS}^2] \quad \textbf{Forward Drain Bias} \tag{766.5B}$$

$$G_{MR} = (Z/L)\mu_n C_o V_{DS}[1 - \tfrac{1}{2}V_{DS}/(V_{GS}-V_{GT})] \quad \textbf{Reverse Drain Bias} \tag{766.6A}$$

$$G_{MF} = (Z/L)\mu_n C_o V_{DS}[1 + \tfrac{1}{2}V_{DS}/(V_{GS}-V_{GT})] \quad \textbf{Forward Drain Bias.} \tag{766.6B}$$

These equations demonstrate the statement just made, namely, I_{DSR}(normal) < I_{SDF}(forward-bias) and G_{MR}(normal) < G_{MF}(forward-bias). Thus, BiMOS built using the forward-bias mode of its MOST will have a higher transconductance than using the normal reverse-bias mode of its MOST.

The above equations are useful in the actual analysis of the BiMOS characteristics. They also demonstrate another point: it is easier to mathematically analyze the BiMOS if we label the emitter island of the BJT the drain of the MOST and short-circuit the source of the MOST to the substrate-base well, because we can use the familiar and simple MOS equation. But, for easy visualizing and deciphering the physics, the emitter island should be called the source since it is the source that supplies or injects the minority carriers into the base layer of the BJT and into the gate-field-induced surface inversion channel of the MOST. The latter, emphasizing the physics, was the key used to decipher the transistor structures disclosed in the three Lilienfeld patents of the 1930's, and the 1948 transistor patents of Bardeen-Brattain and of Shockley, which resulted in the recognition that the Bardeen-Brattain patent had actually disclosed the first junction-gate field-effect transistor (with a surface-field-induced surface channel) in addition to the point-contact transistor. {See sections III.A and III.B of Reference [766.5].}

(3) The two-transistor circuit of Fig.766.1(d) immediately suggests that the short-circuit can be moved to the lower BJT, Q_1, shown in Fig.766.1(e). This gives the new configuration in which the drain/substrate n+/p junction of the nMOST is reverse biased and connected in parallel with the collector-base junction or in series with the base terminal of the n/p/n-BJT. This topological maneuver has given a new circuit-device configuration which will be discussed in the next subsection since it is the very basic element of the BiCMOS invented by H.C. Lin via another circuit evolution route.

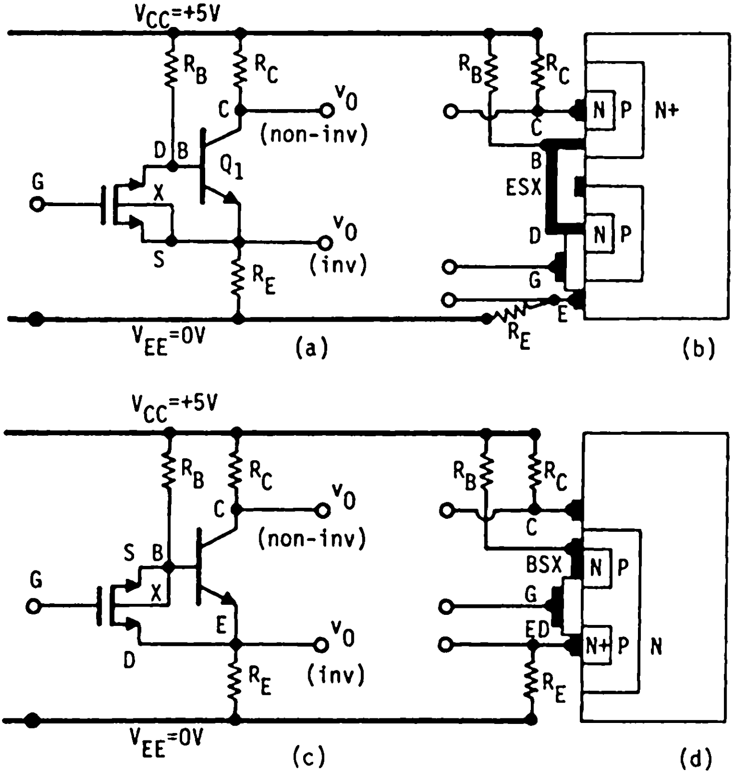

V_CC=+5V
R_B
R_C
C
v_0
(non-inv)
D
B
Q_1
G
X
S
v_0
(inv)
R_E
V_EE=0V
(a)
R_B
R_C
C
N
P
N+
B
ESX
D
G
E
R_E
(b)
V_CC=+5V
R_B
R_C
C
v_0
(non-inv)
S
B
G
X
E
D
v_0
(inv)
R_E
V_EE=0V
(c)
R_B
R_C
C
BSX
N
P
G
ED
N+
P
N
R_E
(d)

Fig.766.2 A comparison of the similarity of circuit and large physical layout differences of the EB-shunt nMOST-n/p/n pair with forward and reverse drain/substrate bias. Reverse drain/substrate bias: (a) circuit diagram and (b) physical layout. Forward drain/substrate bias: (c) circuit diagram and (d) physical layout.

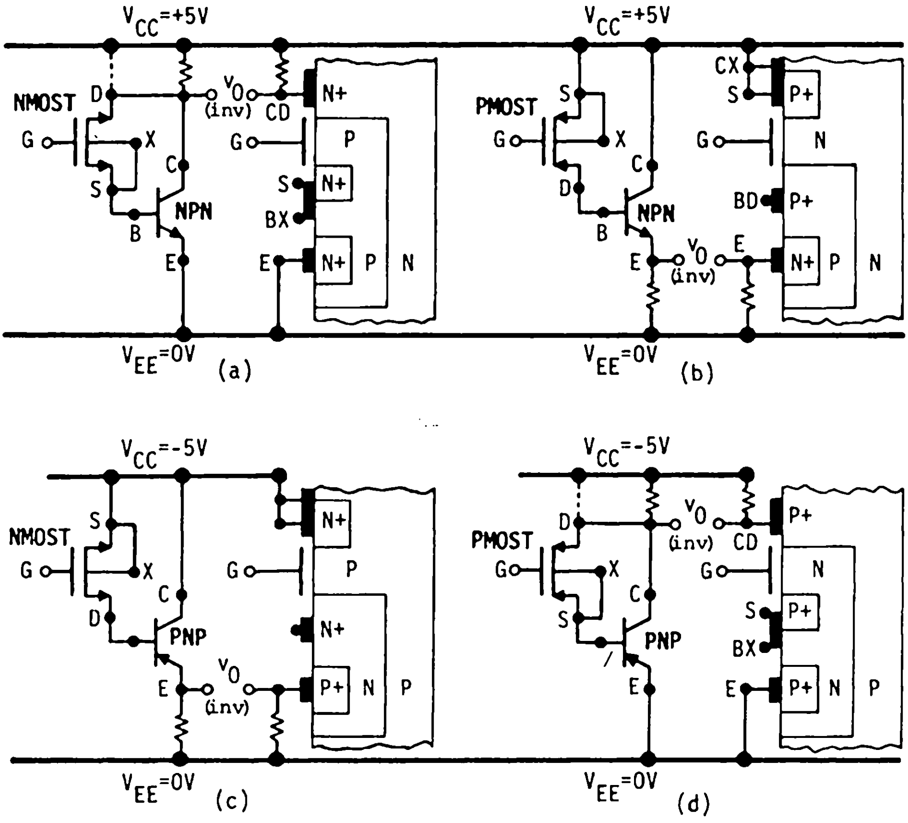
The second group of four configurations of BiMOS with the EB shunted by a MOST consists of the same four combinations of nMOST and pMOST with nBJT and pBJT, except that the drain/substrate junction of the MOST is reverse biased. Without drawing any diagrams [see Figs.766.2(a) and (b) if necessary] one can immediately conclude that physical realization must be more complicated because the emitter/base junction is forward biased while the drain/substrate junction is reverse biased, thus, they cannot share a common p-Si region for the p-side and n-Si region for n-side of their p/n junction. Two separate or isolated p-wells (or n-wells) are required for all four circuits in order to isolate the MOST source/substrate p/n junction completely from the base/emitter p/n junction of the

BJT so as to allow the drain p/n junction to be reverse biased while the base/emitter p/n junction to be forward biased. The complete isolation requires eight different n-to-p connections by metal interconnects over insulators (SiO_2). The eight are: the n-type, p-type, source, emitter, substrate, and base. A representative example is the nMOST-n/p/n circuit shown in Fig.766.2(a) whose physical implementation is shown in Fig.766.2(b). Figures (c) and (d) below (a) and (b) respectively are the corresponding circuits for forward drain bias from Figs.766.1(a) and (b). The (a)-(c) comparison illustrates the very simple circuit difference between forward and reverse biased drain junction, in contrast to the very large difference in physical layout in the (b)-(d) comparison (two wells versus one well). The former (two wells) increases monolithic integration complexity and hence manufacturing cost. The description of the EB-shunt MOST with reverse-biased drain junction is left as an exercise since it operates similarly as the forward-biased drain or source junction just described in connection with Figs.766.1(a) to (d).

The inherent property of the EB-shunt circuit, namely, the MOST is used as a **shunt resistance** or **shunt current path** across the emitter-base junction, is very much like the methodology of current steering by the input voltage used in the ECL circuit described in section 764. The constant current hinders the use of the zero standby power feature of the CMOS even when the drain/substrate junction is reverse biased. Thus, we shall not describe the operation details of the eight BiCMOS circuits using the EB MOST-shunt, four with forward and four with reverse biased drain junction. However, design advantages in layout, larger drive capability, and even speed may still make these EB-shunt BiCMOS current steering circuits useful in certain integrated circuit designs. Indeed they are used in the latest (1990) BiCMOS and CBiCMOS to help pull down and pull up the output in order to attain full swing or rail-to-rail swing ($V_{OH}=V_{HH}$ and $V_{OL}=V_{LL}$ without any voltage loss) which will be described in the next subsection. Furthermore, these shunt circuits appear in several elegant topological symmetry groups and evolution paths when all the possible combinations of the eight MOST-BJT pairs were systematically explored by this author to give the results presented in this section.

Series-BiMOS: MOST in Series with Base Terminal

This family of BiMOS circuits uses the MOST as the **on-off series resistance** or **on-off switch** to supply and interrupt the base current of the output BJT which turns the BJT on and off. It not only amplifies the g_m of the MOST by β_F but also reduces the power dissipation by 50% since the BJT is cut off during half of the switching cycle when the base current is cut off. The preliminary circuit was shown in Fig.766.1(e). The four basic 2-transistor series BiMOS circuits containing an input MOST and an output BJT and their emitter-up physical implementations are shown in Figs.766.3(a) to (d). The drain/substrate junction is reverse biased and the source/substrate junction is zero biased so that the MOST is in the normal operation mode. The key feature is that the MOST is connected **in series** with the base lead, i.e., between the base terminal and the power supply V_{CC}, rather than functioning as a shunt of the collector-base junction to divert a constant current like the EB-shunt BiMOS described in the previous subsections. Thus, the MOST completely shuts down both itself and the BJT because the base current is also turned off when the MOST is turned off by an input gate pulse. Only leakage currents will flow through the transistors. Nevertheless, in the other half of a switching cycle, both the MOST and the BJT draw a high current when the BJT is turned on by the conducting MOST which was turned on by a gate pulse.


V_{CC}=+5V
NMOST
D
G
X
S
C
B
NPN
E
V_O (inv)
CD
N+
P
S
N+
BX
N+
P
N
V_{EE}=0V
(a)
V_{CC}=+5V
PMOST
S
G
X
D
C
B
NPN
E
V_O (inv)
CX
S
P+
N
BD
P+
E
N+
P
N
V_{EE}=0V
(b)
V_{CC}=-5V
NMOST
S
G
X
D
C
PNP
E
V_O (inv)
N+
P
N+
P+
N
P
V_{EE}=0V
(c)
V_{CC}=-5V
PMOST
D
G
X
S
C
PNP
E
V_O (inv)
CD
P+
N
S
P+
BX
E
P+
N
P
V_{EE}=0V
(d)

Figs.766.3 The four basic building blocks of series-BiMOS circuits using the MOST in series with the base to control the base current. (a) NBi-NMOST (b) NBi-PMOST (c) PBi-NMOST and (d) PBi-PMOST.
)

BiCMOS and CBiCMOS Inverters

The large standby current during half of the switching cycle in the four basic 2-transistor series-BiMOS circuits just described in Fig.766.3(a)-(d) can be eliminated if two of the four circuits are connected in series so that the current is cut off in both halves of the switching cycle. There are $4^2=16$ combinations, but the combinations that should come to mind immediately are those which give the CMOS configuration because we learned in section 672 that CMOS dissipates no or little standby power. This eliminates all but four of the sixteen circuits; two are **BiCMOS (Bipolar CMOS)** and two are **CBiCMOS (Complementary Bipolar CMOS)**. They are shown by the vertical 2-T pairs in Figs.766.4(a)-(d).

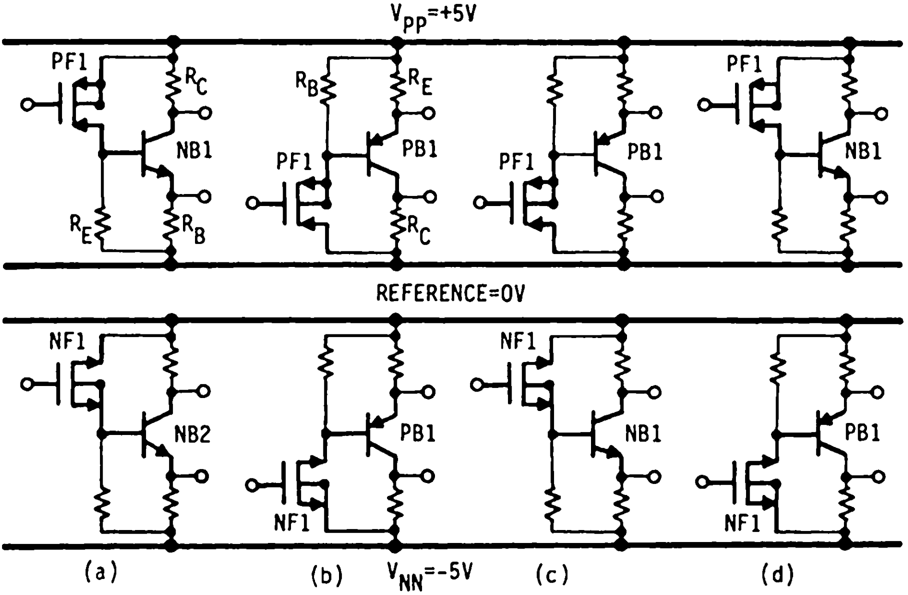
The upper and lower parts of each vertical 2-T pair can then be merged. In this merger, the R_C's and R_E's are short-circuited since the BiMOS connects the output directly to +5V or -5V without the need of a load resistance. The R_B's are opened since the MOST controls the base current. The resulting four inverter circuits are shown in Figs.766.5(a)-(d) between the two railway tracks or d.c. power supply buses. A ground bus is not necessary but is available and not shown. Later, we shall show that some of the R's should be retained to give full-swing or rail-to-rail output voltage swing.

Figures 766.5(a) and (b) are the two BiCMOS inverter circuits. (a) is the **NBiCMOS** using two n/p/n transistors in the output which is the historical first (invented by H.C.Lin in 1969 [766.6]) and the fastest because n/p/n is faster than p/n/p. (b) is the **PBiCMOS** using two p/n/p transistors. Figures 766.5(c) and (d) are the two **CBiCMOS (Complementary Bipolar CMOS)** inverters. (c) is the CBiCMOS with **collector output** and (d) is the CBiCMOS with **emitter output.**

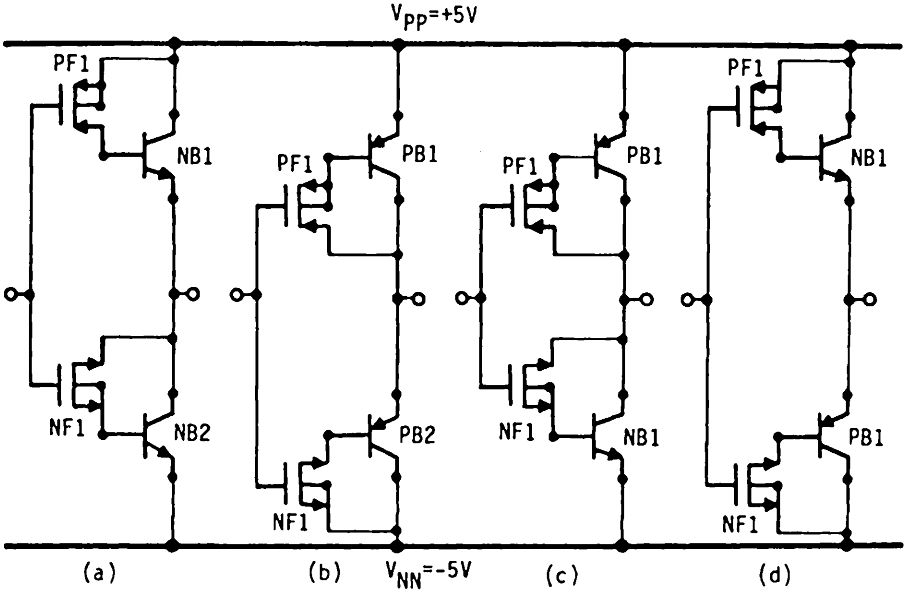
[766.6] Hung Chang Lin, J.C.Ho, Ramachandra R. Iyer, and Ken Kwong, "Complementary MOS-Bipolar transistor structure," IEEE Transaction on Electron Devices, ED-16(11), pp.945-951, November 1969. The latest circuit developments are reported in [766.7] on BiCMOS and [766.8] on CBiCMOS. And the latest technology and circuit applications are reported in the five annual conference proceedings: IEDM, ISSCC, BCTM (Bipolar Circuits and Technology Meeting), and US-Japan Symposiums on VLSI Technology and Circuits.

[766.7] Antonio R. Alvarez (editor), **BiCMOS Technology and Applications**, Kluwer Academic Publishers, Boston, 1989. Chapter 1 by the editor gives a survey and projection. Chapter 5 by K. Deierling reviews the digital circuits.

[766.8] Hyun J. Shin, Chih-L Chen, Eric D.Johnson and Yuan Taur (IBM) "Full-swing complementary BiCMOS logic circuits," Proceedings of the 1989 IEEE Bipolar Circuits and Technology Meeting, pp.229-232, September 18-19, 1989; and Hyun J. Shin, "Full-swing logic circuits in a complementary BiCMOS technology," 1990 Symposium on VLSI Circuits, Digest of Technical Papers, pp.89-90, June 7-9, 1990. IEEE Cat. No.90 CH2885-2.


V_{PP}=+5V
PF1
R_C
NB1
R_E
R_B
R_B
R_E
PF1
PB1
R_C
PF1
PB1
PF1
NB1
REFERENCE=0V
NF1
NB2
PB1
NF1
NF1
NB1
PB1
NF1
(a)
(b)
V_{NN}=-5V
(c)
(d)

Figs.766.4 The four bipolar-MOS inverters from combining the four 2-T BiMOS with nearly zero standby power dissipation. **(a)** NBiCMOS. **(b)** PBiCMOS. **(c)** CBiCMOS. **(d)** CBiCMOS.


V_{PP}=+5V
PF1
NB1
NF1
NB2
PF1
PB1
PB2
NF1
PF1
PB1
NF1
NB1
PF1
NB1
PB1
NF1
(a)
(b)
V_{NN}=-5V
(c)
(d)

Figs.766.5 The **merged** four inverters of Figs.766.4. **(a)** NBiCMOS. **(b)** PBiCMOS. **(c)** CBiCMOS. **(d)** CBiCMOS.

The four inverters have two unique common features which were the keys used for selecting the four 2-T BiMOS inverter circuits to form the CMOS input of each of the four 4-T BiCMOS and CBiCMOS inverter circuits shown in Figs.766.4(a)-(d) and Figs.766.5(a)-(d). The keys are: (i) a pMOST whose p-source/n-substrate is connected directly or indirectly (via the EB junction of the BJT) to the +5V power supply and (ii) an nMOST whose n-source/p-substrate is connected directly or indirectly to the -5V (or 0V) power supply. The four 4-T BiCMOS-CBiCMOS inverters are distinguished by four different 2-BJT output driver configurations of the totem-pole, push-pull or pull-pull type: one of the two BiCMOS inverters uses two n/p/n BJTs to be called NBiCMOS and the other uses two p/n/p BJTs to be called PBiCMOS. The two CBiCMOS inverters uses one n/p/n and one p/n/p to give a complementary or push-pull bipolar output driver.

Figures 766.6(a) to (d) on the next page show the one possible physical implementation of each of the two BiCMOS and two CBiCMOS inverters using emitter-up BJTs. Using emitter-down or a combination of emitter-up and emitter-down BJTs gives three additional physical realizations for each inverter.

The operation of these BiCMOS and CBiCMOS circuits can readily be demonstrated. Consider the NBiCMOS in Fig.766.5(a). When input is +5V or HI, the pMOST PF1 is off so the n/p/n NB1 is also off. The nMOST NF1 is on which sources the base current of NB2 (i.e. turns NB2 on but not passing a d.c. current unless the output or input is changing with time) to pull the output down towards -5V. But due to the emitter-base forward bias of NB2, the output magnitude is decreased to $V_{OL}=V_{NN}+V_{BE}=-5+0.8=-4.2V$. This also clamps NB2, preventing it from getting into saturation thereby improving turn-off time.

Similarly, when the input is -5V or LO, N_{F1} and N_{B2} are turned off, and PF1 and NB1 are turned on and they pull the output up towards +5V. Base-emitter voltage reduces the output to $V_{OH} = V_{PP} - V_{BE\text{-}NB1} = +5 - 0.8 = +4.2V$ at low load current. At high load current, $V_{OH} = V_{PP} - V_{CE\text{-}sat\text{-}NB1} = +5 - 0.1 = +4.9V$. The standby or quiescent current is nearly zero, just the leakage current of the reverse biased p/n junctions.

The transconductance is the MOST transconductance amplified by β_F or $G_m=\beta_F g_m$. The output impedance for either the high or low output state is the sum of the MOST drain resistance and the BJT base-emitter diode resistance reduced by the forward beta. Let $\beta_F=100$, $I_L=1mA$, $W/L=5$, $\mu_n=\mu_p=250cm^2/V\text{-}s$, $X_o=345A$, then $C_o = \varepsilon_o/X_o = 3.9x8.854x10^{-14}/345x10^{-8} = 10^{-7}F/cm^2$, $g_m = g_d = (W/L)\mu_n C_o(V_G-V_{BE}) = 5x250x10^{-7}x(5-0.8) \approx 0.5mS$. Thus, $G_m = \beta_F g_m = 100x0.5mS = 50mS$, $r_{be} = (kT/qI_E) = (25mV/1mA) = 25\Omega$ and $Z_o = (r_d+r_{be})/\beta_F = (2000+50)/100 = 20\Omega$. Usually $\mu_n > \mu_p$ and the two MOSTs can be made nearly identical by making the aspect ratios and oxide thicknesses of the two MOSTs satisfy $(W_p/L_p)\mu_p C_{op} = (W_n/L_n)\mu_n C_{on}$.

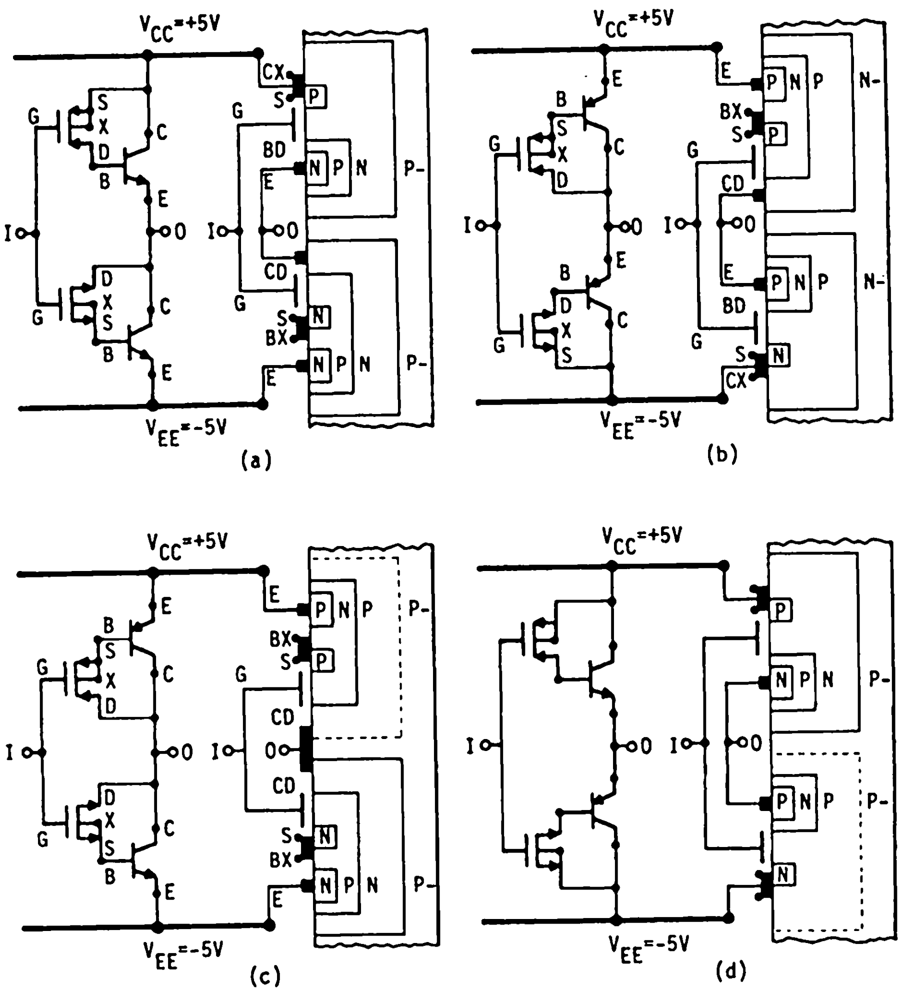


Fig.766.6 Physical realization of the 4-T BiCMOS and CBiCMOS inverters. (a) NBiCMOS. (b) PBiCMOS. (c) CBiCMOS with collector output. (d) CBiCMOS with emitter follower output.

The preceding numerical example shows that the BiCMOS and CBiCMOS inverters have all the desirable features of an ideal logic inverter: high (nearly infinite) input impedance of a small input capacitance ($1\text{fF}/\mu\text{m}^2$ = 10 to 100fF), low output impedance ($\approx 20\Omega$), and high transconductance (50mS), which would give an intrinsic speed of C/g_m = 50fF/50mS = 1ps and which is capable of driving a large capacitance load without slowing down, $C_L/g_m = 1\text{pF}/50\text{mS} = 20\text{ps}$.

However, there are two performance degradation factors. (i) The CE BJT turn-off delay is long because its base lead is opened by the turned-off series MOST in the base lead. This limits the speed of removing the stored minority carriers in the base layer, whether the BJT's are driven into saturation or not (in these BiCMOS circuits, they are not but there is still a stored base charge to be extracted), to the slow recombination mechanism at the recombination centers in the base as indicated by (753.10) whose time constant is the recombination lifetime given by (755.5), $t_5 = \tau_B = \beta_F t_B = \beta_F K_5(X_B{}^2/2D_B)$. It was shown in (755.17A) and (755.17B) that the shortest recombination lifetime is still $>$1ns when Si is doped to the highest practical gold concentration. It was shown after (753.10) that a much faster turn-off is to short-circuit the base-emitter and base-collector junctions whose turn-off time constant is $t_B = \tau_B/\beta_F \propto$ 1-10ps. (ii) The output voltage swing is reduced, as just illustrated by the numerical example, from the full swing or rail-to-rail value of -5V to +5V to -4.2V to +4.2V. The lower output at each inverter stage could make a long-chain of inverters inoperative because the voltage swing near the end of the chain could be less than the threshold voltage of the next input CMOS. The lower output also reduces the noise margin (noise from the coupling of random digital signals passing through closely spaced transistors and interconnect transmission lines) because of process variation and operational drift of the MOST threshold voltage on a chip containing millions of MOSTs. The loss of output voltage swing increasingly limits performance as the dimension is reduced to below 0.5-μm because the power supply voltage must also be reduced, to 3.3V or less, in order to reduce the threshold voltage drift due to hot electrons in high electric fields.

These two performance degradation factors are being overcome in the latest development efforts (1989-1991-onwards). To draw out or extract the stored base minority carriers via the base lead, **switched** on-off resistances (or EB-shunt MOSTs described in the preceding subsection) are used to shunt the EB terminal, known as **active pull-up** and **active pull-down**. To restore the output to full swing or rail-to-rail, **unswitched** or 'permanent' resistances (gate-drain-tied EMOSTs or DMOSTs) are used to shunt the EB, known as **passive pull-up** and **passive pull-down**, but these permanent resistances can increase the power dissipation. Alternatively or simultaneously, a CMOS output (which provides the full-swing) is added in parallel with the BJT output to pull-up and pull-down the remaining voltage losses at reduced speed. It is obvious from this description, very many circuits, using MOSTs only in EB-shunts, CE-shunts and output-shunts, can reduce and eliminate these two performance degradation factors. Then cost or real-estate and yield become the prime engineering-economic factors for selection in a particular application. We shall only describe a few simple examples in the following paragraphs to illustrate the principle of eliminating these two limitations.

EB-shunt resistance gives the two simplest passive pull-up, pull-down, and speed-up scheme. The two circuits are shown in Figure 766.7(a) and (b) for the NBiCMOS which was first pedagogically described by K. Deierling [766.7].

Current and voltage labels are added in our figures (a) and (b) to be used in a description of switching operation to follow. These two circuits use a EB-shunt resistance in parallel with the EB diode of each of the two BJTs. They are labeled R_1 and R_2. Notice that figure (a) is a totem-pole of a PMOST-n/p/n BiMOS sitting on top of a NMOST-n/p/n BiMOS, which is the original circuit-device evolution path traced by H.C.Lin in his invention of the BiCMOS [766.6]. On the other hand, figure (b) has a true CMOS input in accordance with the original spirit and intent of CMOS-BJT. It is one resultant circuit configuration of the many 4-T circuits following this second circuit evolution path that gave the four without shunts shown in Figs.766.5(a)-(d). Thus, these two circuits demonstrate two circuit evolution paths to reach ideal performance characteristics at the expense of increasing complexity and transistor counts.

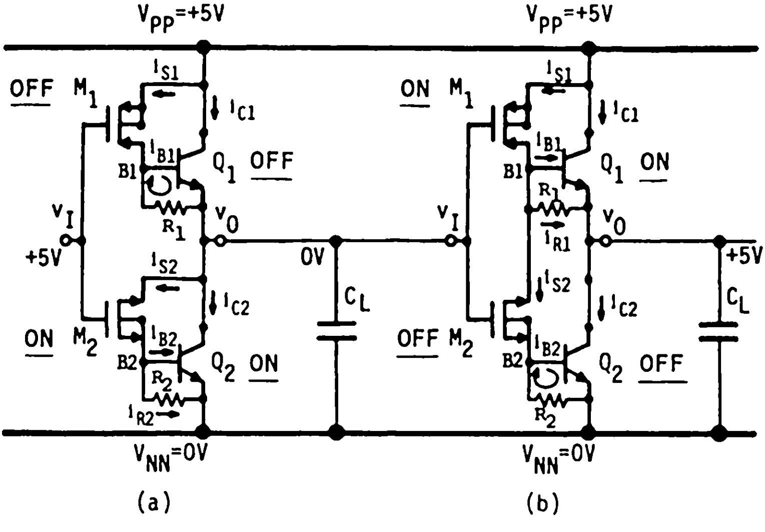


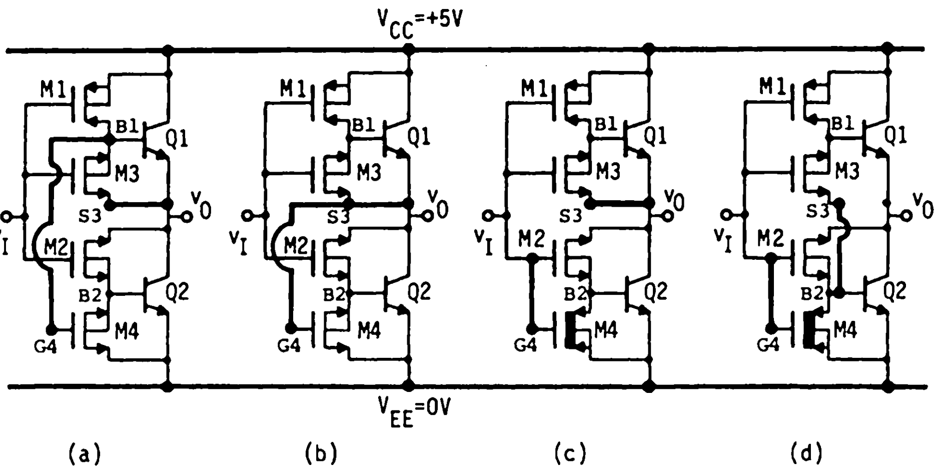
Fig.766.7 The two simplest NBiCMOS circuits that use EB-shunt resistance for **rail-to-rail full-swing** pull-up, pull-down, and for speed-up. **(a)** Totem-pole nBJTpMOST sitting on top of nBJTnMOST. **(b)** True CMOS input. Note, the resistances pull the output to ground or full V_{PP} to give the rail-to-rail full-swing.

Let us now analyze the operation. Consider Fig.766.7(a) which is labeled for input=HIGH=+5V. The upper part of the totem-pole, E-pMOST (M1), is turned off, $i_{S1}=0$, and the upper n/p/n BJT (Q1) is also turned off or turning off, $i_C(t)\rightarrow 0$, since its base current source ($i_{S1}=0$ through M1) has been cut off. The minority carrier (electrons) charges, stored in the n/p/n (Q_1) from the previous half cycle when Q_1 was on (either in or outside of saturation depending on r_{d1}, R_1, r_{be1} and β_{F1}), are now extracted out by the reverse-biased collector-base junction and the short-circuited-through-R_1 base-emitter junction. Thus, turn-off of Q_1 is sped up towards the time constant $t_{B1} = \tau_{B1}/\beta_{F1} = K_5 X_{B1}^2/2D_{B1} \approx$ 1-10ps. The lower part of the totem-pole, E-nMOST (M2), is turned on which sources the base current of the lower n/p/n BJT (Q2). Any voltage appearing at the output $v_O(t)$, such as the +5V of the previous half cycle of the switching waveform on a load capacitance C_L between the output node and ground, will be discharged (i) through Q_2 to a LO voltage of $V_{CE\text{-}sat2}$ ($\approx$+0.1V), and (ii) simultaneously through M_2 in series with the parallel combination of R_2 and $r_{be\text{-}2}$ to a voltage of $V_{BE\text{-}sat2}$ without R_2 and to ground with R_2. Thus, R_2 pulls down the output to ground (0V) or the lower rail potential (if it was at -5V instead of 0V). The output resistance that discharges the load capacitance is given by $Z_o \approx \{r_{d2}+[R_2 r_{be2}/(R_2+r_{be})]\}/\beta_F = \{2000+[2000\text{x}50]/(2000+50)\}/100 = 20.5\Omega$ from using the numerics given before Fig.766.6 and assuming R_2 is equal to the channel resistance of the nMOST (M2). For a 100fF=0.1pF load (rather larger than sub-micron gate in a FAN-OUT=1), the discharging time constant is $20.5\Omega \cdot 100\text{fF} = 2.05\text{ps}$. This is comparable with the turn-off delay of Q_1 just described and the intrinsic delays of the MOSTs and the BJTs. Thus, all the circuit and device delays must be considered to give the total delay and waveforms which is studied analytically and numerically in the next circuit course and analyzed in the trade using SPICE, a transient circuit-device analysis program.

Next consider input=LOW=0V shown in Fig.766.7(b) which is drawn following figure (a) as if the two inverters are cascaded! The lower part of the CMOS, the nMOST (M2) is turned off and the stored base charge in Q_2 is extracted out via the EB-shunt R_2. The upper part of the CMOS, the pMOST (M1) is turned on which sources the base current for Q_1, $i_{B1}(t)$, and the current through resistance R_1, $i_{R1}(t)$. The output voltage is pulled up by Q_1 to $V_{PP}\text{-}V_{CE1} = 5\text{-}0.1 \approx +4.9\text{V}$ and R_1 pulls the output simultaneously to the upper rail voltage, $V_{PP} = +5\text{V}$. The delay in charging up a load capacitance, C_L, from its previous state of 0V by the output voltage to +5V can be similarly estimated like that for figure (a).

The two resistors can be replaced by two MOSTs, either forward or reverse biased. The MOSTs can be switched (i) on and off to minimize the increase of power dissipation but losing full-swing/rail-to-rail capability, (ii) to a high and low value to maintain full-swing while minimizing delay and additional power dissipation, and (iii) not switched, like the fixed resistances just described in Figs.766.7(a) and (b).

Figures 766.8(a)-(d) show four NBiCMOS inverters all using on-off switched forward-biased EB-shunt (M3 and M4) across the two n/p/n BJTs (Q1 and Q2). Since the shunt is switched off, rail-to-rail full-swing is not attained. Only turn-off speed is improved. The difference among this four are in the lower-part of the totem-pole. Figures 766.8(a) and (b) use enhancement-mode forward-biased nMOST EB-shunt but (a) uses the input to turn on and off the nMOST EB-shunt while (b) uses output feedback to turn on and off the nMOST EB-shunt. Figures 766.8(c) and (d) use depletion-mode forward-biased pMOST EB-shunt, and both use the input to turn on and off the pMOST EB-shunt. Additional turn-off speed-up is obtained in Fig.766.8(d) by tieing the drain of the EB-shunt (M_3) of Q_1 to the base (B_2) of the lower n/p/n Q_2. Figure (d) is the original modified BiCMOS circuit discovered by H.C.Lin given in Fig.10 of [766.6].



Figs.766.8 Four NBiCMOS inverters using on-off switched and forward-biased MOST for the EB-shunts to speed up the BJT turn-off transient without attaining rail-to-rail full-swing. Lower nMOST EB-shunt pull-down (a) without output feedback and (b) with output feedback. Lower pMOST EB-shunt pull-down (c) standard and (d) with additional speed-up.

In the following several paragraphs we shall describe in qualitative detail several modifications of the two basic or bare CBiCMOS inverters shown in Fig.766.5(c) and (d) because these are likely to dominate all future MOS and MOS-Bipolar integrated circuits. To be described are the three modifications-additions that will (i) speed up the BJT part of the circuit to approach the intrinsic delay, t_B, (ii) give rail-to-rail full-swing output, and (iii) provide protection against transients from the load. It will be demonstrated that the addition of one element (a E-MOST parabolic-resistance/Shockley-diode) connected in the appropriate way in the EB-shunt position of each stack of the 2-stack totem-pole of the four basic-bare CBiCMOS inverter, shown in Figs.766.5(c) and (d), will give all of these features. Nevertheless, current engineering development reports (1989-1990) [766.8]-[766.10] have dealt exclusively with the emitter-follower CBiCMOS, shown in Fig.766.5(d), and have proposed circuits with a substantial number of additional MOSTs in order to give these three features [766.8]. We shall use the emitter-follower to introduce the description and then give details for the collector output modifications.

[766.9] A.Watanabe, T.Nagano, S.Shukuri, T.Ikeda (Hitachi), "Future BiCMOS technology for scaled supply voltage," Technical Digest of IEDM-89,pp.429-432, IEEE Catalog No.89CH2637-7. The emitter-follower CBiCMOS is the only one described and shown in Fig.10(b) on p.432.

[766.10] M.Fujishima,K.Asada,T.Sugano (U.Tokyo), "Appraisal of BiCMOS from circuit voltage and delay time," Digest of Technical Papers, 1990 Symposium on VLSI Circuits, pp.91-92, IEEE Catalog No.90CH2885-2. Again the emitter-follower CBiCMOS is the only one described and shown in Fig.1(b) on p.92.

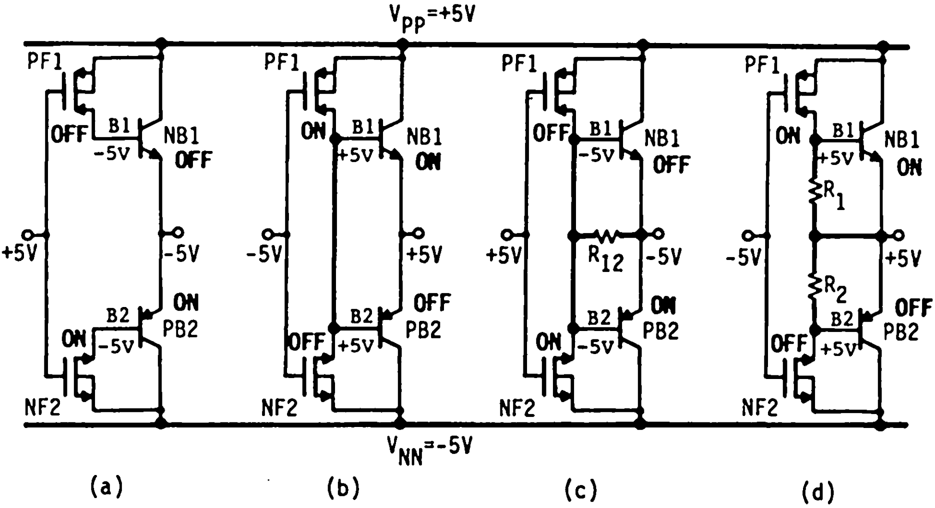


Fig.766.9 Evolution of modification of the basic-bare emitter-follower CBiCMOS inverter with EB-shunt resistors. (a) Original basic-bare circuit. (b) Tied base nodes. (c) Shared EB-shunt resistor. (d) Separate EB-shunt resistors.

The simplest scheme to give the three characteristics is to use a resistance EB-shunt whose pedagogical evolution sequence is shown in Figs.766.9(a)-(d) for the emitter-follower CBiCMOS inverter. Figure (a) is the bare basic circuit, initially shown in Fig.766.5(d) which has all three deficiencies repeated as follows. (i) The output response to an input switching step from +5V to -5V is slow because the p/n/p BJT (PB2) is turned off by opening its base via turning off the nMOST (NF_2). Opening the base means that the stored base charge must be extracted by recombination with a time constant equal to the recombination lifetime of τ_B which is $\beta_F(\approx 100)$ larger than the base diffusion transit time, $t_B = K_5 X_B^2/2D_B = \tau_B/\beta_F$. (ii) The output does not achieve rail-to-rail full swing because of the EB voltage drops, V_{BE1} and V_{BE2}. (iii) And, the EB diodes are also not protected against load transients (±10V or >|±5V|) that would bias the EB junction into breakdown and damage the EB junction. For example, a load voltage transient pulse of $v_O(t)=-10V$ amplitude in figure (a) will reverse bias the EB junction of the n/p/n to -5V momentarily which may cause excessive current and breakdown. In future picosecond, 100GHz, BiCMOS circuits, the breakdown (or tunnel) voltage could drop to 2V and the power supply voltage of 1.5V.

The first pedagogical step is the addition of a wire to tie the two base nodes as shown in Fig.766.9(b). The successfully implemented-characteristics among the three are numbered with underlined numerals. This tie line (**tied base**) makes the CMOS and the CBJT connected in truly isolated cascade without any feedforward or feedback links between the two cascaded stages. It is actually the starting point of an alternative circuit evolution path. The effects on circuit performance are as follows. (i) Speed-up is achieved since PF1 will discharge PB2 and PF2 will discharge PB1. (ii) Rail-to-rail full-swing is not achieved because the forward emitter-base voltage drop is still there. (iii) Output protection is achieved from cross paralleling the two EB p/n junctions: they clamp each other, preventing any large transient load voltages from reverse biasing the EB junctions into avalanche breakdown which would degrade the BJTs.

Addition of one **shared EB-shunt** resistance R_{12} shown in figure (c) improve speed and also gives the other characteristic: (i) additional speedup of the turn-off transient of the p/n/p BJT (PB_2) from $\tau_B=\beta_F t_B$ to t_B when v_I is switched from +5V to -5V, and (ii) rail-to-rail full-swing because the resistance R_{12} has no built-in potential drop so it shunts out the forward voltage drop or built-in voltage of the EB p/n junctions. This is known as the **passive** pullup and pulldown scheme since the resistance R_{12} is passive or constant and is not switched.

The successful operation of circuit (c) to attain all three characteristics just described has assumed identical characteristics of the two MOSTs and two BJTs so that the interaction of the switching transients in the two totem-poles are completely synchronized with identical waveforms and delays. A difference between the nMOST and pMOST or n/p/n and p/n/p BJTs could create a large transient current during switching, for example, when the turn-off of the nMOST is much faster than the turn-on of the pMOST. But it is practically impossible to make the two MOSTs identical and the two BJTs identical. Firstly, identical characteristics under all operating conditions cannot be attained readily via device design to compensate for differences in: electron and hole mobilities and lifetimes, threshold and built-in voltages, impurity doping profiles, and temperature variations of these parameters. Secondly, identical characteristics is not tenable due to manufacturing process variations across the chip area.

However, improvement of circuit (c) against these variations can be readily made, as indicated in figure (d), by splitting the resistance R_{12} into two, R_1 and

R_2, in series with the two base nodes. Then, each stack of the totem-pole has its own current path so that the turn-on/turn-off currents in the two BJTs do not interact through a common resistance. Thus, slight differences in waveforms and delays due to transistor parameter variations in the two stacks of the totem-pole can be tolerated and will only lower the circuit speed or clock slightly.

In addition, the EB junctions are no longer reverse biased during the switching transient in circuit (d). This is desirable even if the reverse bias on one EB junction in circuit (c) is clamped by the other EB junction which is forward biased to about 1V. The clamping action against large load voltage transients or noise (such as ±10V) still exists in circuit (d) but it is slightly less effective due to the series R_1 and R_2. This inverter, (d), was recently (1990) analyzed by Professor Sugano and his associates at the University of Tokyo [766.10].

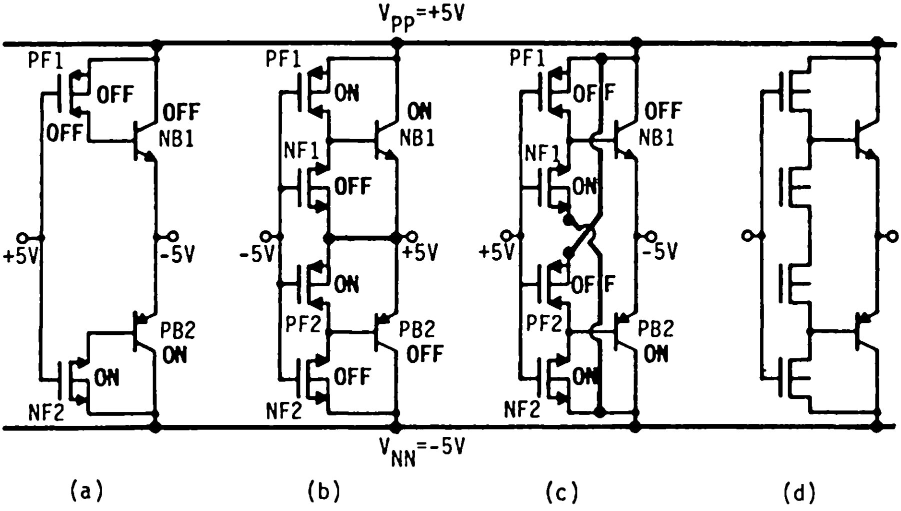


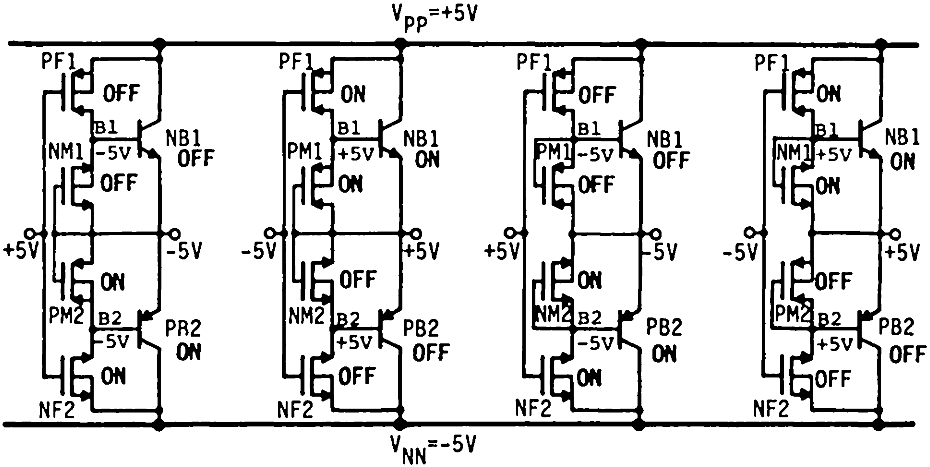
Fig.766.10 One evolution path of using CMOS as active or switched EB-shunt in emitter-follower-output CBiCMOS inverter. (a) Original basic-bare circuit. (b) CMOS EB-shunts with feedback via the source. (c) CMOS EB-shunts without feedback. (d) For exercise use.

It is more difficult to physically realize the resistance shunts shown in Figs.766.9(c)-(d) in monolithic integration than to use diodes, transistors, and diffused resistors or E-MOST or D-MOST resistors. Figures 766.10(b)-(c) show the two earliest (1989-1990) variations of using switched MOSTs with reverse-biased drain/substrate junction as the EB resistance shunts in the emitter-follower CBiCMOS inverter of (a) or Fig.766.5(d). These are known as active EB-shunts since the MOSTs of the EB-shunt are switched. Circuit (b) has the active EB-shunt in the feedback loop (source tied to the output) and circuit (c) does not (source tied to the power supply). Unfortunately, these two inverters still have some of the three deficiencies. (i) Rail-to-rail full-swing is not attained because the EB-shunt resistance is switched off when the associated base-current-biasing MOST and the BJT are switched on, leaving still an EB built-in voltage short-fall. (iib) Inverter in figure (b) is slow in switching because the source of the EB-shunts are connected to the output whose switching delay prevents the EB-shunts from turning on instantaneously as the input voltage switches between the two states. For example, the nMOST (NF1) EB-shunt will not turn on at once when the input voltage switches from -5V (the quiescent value shown in the figure) to +5V because the source of the EB-shunt MOST is connected to the output, which is still at +5V and takes time to discharge to -5V. This output delay then delays the turn-off of the n/p/n (NB1). (iic) Inverter in figure (c) overcomes this output feedback delay by connecting the source of each EB-shunt MOST to the appropriate forward power supply voltage instead of the output. (iii) Protection of the EB junction of the BJTs against load noise voltages which might reverse bias these low-breakdown voltage EB junctions is attained by the MOST EB-shunts at one forward-biased drain/substrate junction voltage $|V_{DX}| \approx 0.8V$ in figure (b) (iiib), but at two series forward-biased p/n junctions from the series drain/substrate and E/B p/n junctions, $|V_{DX}|+|V_{EB}| \approx 1.6V$, in figure (c) (iiic) which is less desirable. (iv) Owing to two CMOS' driven by the input, these two inverters have twice the input capacitance load. The inverter shown in Fig.766.10(c) was first analyzed in 1989 by Hitachi engineers [766.9].

To overcome these deficiencies, inherent in the emitter-follower CBiCMOS with active EB-shunt just described, the IBM group [766.8] has proposed to add two MOSTs to attain each of the three characteristics: (i) speed up, (ii) rail-to-rail full-swing, and (iii) clamp protection. This approach would raise the total transistor/diode count to ten for each inverter. Additional 2-T CMOS was added in the feedback loop to increase the speed further. Such a large number of transistors per inverter stage increases real-estate, circuit complexity, and noise sensitivity, all of which are undesirable in highest gate density (gate per chip) applications.

The Optimum 6-Transistor CBiCMOS Inverter

A simpler and straightforward solution that gives the three properties all at once was discovered in the present systematic exploration. It is to use the two-terminal E-MOST or D-MOST **diode** as the **nonlinear passive** EB-shunt. This MOST has $V_D=V_G$ and $V_S=V_X$ which provides the two terminals or nodes. It is a **parabolic resistive diode** when the drain/substrate-source junction, V_{DX}, is reverse biased and it is a **forward-biased p/n junction diode** when V_{DX} is forward biased. Thus, if properly connected, the parabolic resistive diode (which has no built-in voltage) would give rail-to-rail full-swing and would also speed up the discharge of the stored base charge rapidly, and the forward-biased p/n junction diode would give the clamp across the EB junctions against load voltage transient that would have reverse biased the EB junctions into breakdown. The characteristics of such an nMOST diode for EB-shunt are: $I_D(\text{reverse bias}) = (W/2L)\mu C_o(V_{DS}-V_{GSX})^2$ and $I_D(\text{forward bias}) = -I_1[\exp(q|-V_{DX}|/kT)-1]$.



(a) (b) (c) (d)

Fig.766.11 The four emitter-follower-output CBiCMOS inverters with EB-shunt using passive unswitched E-MOST diodes. (a) and (c) the drain-junction is forward biased, and (b) and (d), the drain-junction is reverse biased.

There are obviously four circuit configurations from placing the two MOSTs (nMOST and pMOST) at the EB-shunt position of the two BJT's. The four inverter circuits are shown in Figs.766.11(a)-(d). The principle difference is that the drain junction of EB-shunt MOSTs is either forward biased or off in circuits (a) and (c), and reverse biased or off in (b) and (d). The forward-bias configuration is responsible for having the three deficiencies. Thus, inverters (a) and (c) do not have rail-to-rail full-swing output and clamp-protection against reverse biasing the EB junction by large load voltage noise (±10V); and are slow due to both gain loss from the forward-biased EB-shunt and charge stored in the base region of the forward-biased MOST EB-shunt. The inverter circuits (b) and (d) have eliminated all three deficiencies with a pair of EB-shunts that are connected to give reverse bias to the drain/substrate junction. They have (i) fast turn-off and turn-on with little gain loss and switching speed reduction, (ii) rail-to-rail full-swing from the channel resistance of the MOST EB-shunt which has no built-in voltage, and (iii) protection of the EB junction from being reverse biased into breakdown by load voltage noise because it is clamped to the built-in voltage of the forward-bias drain junction.

Four collector-output CBiCMOS inverters using fixed resistances, switched resistances, and the MOST diodes as the EB-shunt to improve the performance are shown in Fig.766.12(a)-(d). Figure (a) is the standard resistance EB-shunt. Figure (b) uses a D-nMOST and a D-pMOST as the EB-shunt which is harder to build and whose input load is twice as high since the gate of four MOSTs are connected to the input. Figures (c) and (d) uses E-MOST with the drain/substrate junction reverse biased to give the parabolic resistance characteristics and the forward-biased drain/substrate junction clamp for protection against power supply bounce. Since the low-breakdown-voltage EB junction is not connected to the output node, output protection using forward-biased p/n junction diode clamp is not necessary provided the load voltage noise is smaller than the fairly high breakdown voltage of the C/B and D/X junctions.

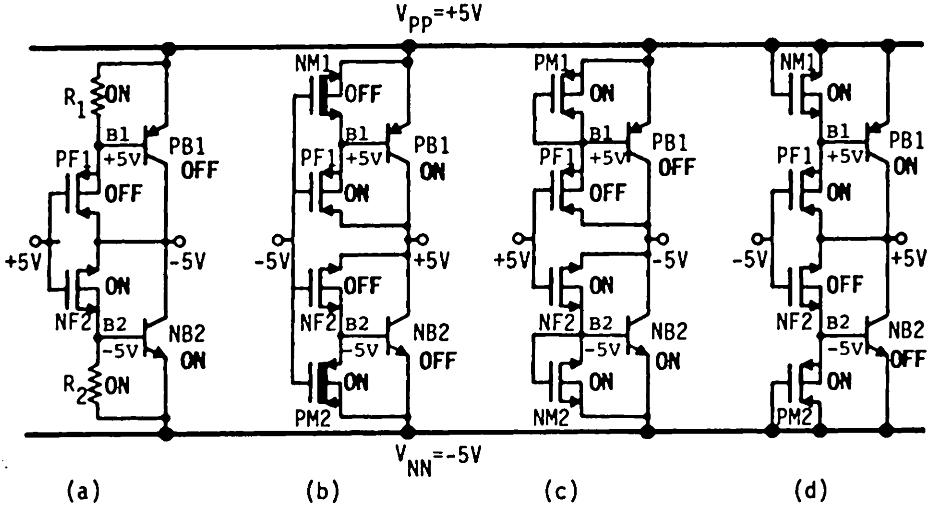

V_{PP}=+5V
R_1 ON
B1
PF1
+5V
PB1
OFF
OFF
+5V
ON
-5V
NF2
B2
-5V
NB2
ON
R_2 ON
NM1
OFF
B1
PF1
+5V
PB1
ON
ON
-5V
OFF
+5V
NF2
B2
-5V
NB2
OFF
ON
PM2
PM1
ON
B1
PF1
+5V
PB1
OFF
OFF
+5V
ON
-5V
NF2
B2
-5V
NB2
ON
ON
NM2
NM1
ON
B1
PF1
+5V
PB1
ON
ON
-5V
OFF
+5V
NF2
B2
-5V
NB2
OFF
ON
PM2
V_{NN}=-5V
(a)
(b)
(c)
(d)

Fig.766.12 The four collector-output CBiCMOS inverters with EB-shunt. (a) Unswitched-passive resistance shunt. (b) Switched-active D-MOST shunt. (c)-(d) Reverse-biased drain-junction shunt.

Thus, the two collector-output CBiCMOS inverters in Figs.766.12 (c) and (d) and the two emitter-follower-output CBiCMOS inverters in Figs.766.11(b) and (d) are the four CBiCMOS inverters that give the best performance. The one difference between the collector and emitter-follower outputs is the minimum input voltage to switch the output state (or voltage level) between HIGH and LOW. In the emitter-follower-output CBiCMOS circuit, the minimum input voltage is the threshold voltage of the MOSTs in the CMOS whose substrate/source are tied to the power supplies. But in the collector-output CBiCMOS inverters, it is a higher voltage consisting of the sum of the threshold voltage of the MOST and the emitter-base built-in voltage of the BJT, because the EB junction of the BJTs is connected in series with the CMOS source/substrate terminal and the power supply. The larger minimum input voltage to switch the output state would limit the application of the collector-output CBiCMOS in low-voltage equipment and submicron devices where low power supply voltage and operating voltage are important operating factors. **Thus, the emitter-follower-output CBiCMOS given by Figs.766.11(b) and (d) are the preferred CBiCMOS inverter circuits.** Nevertheless, technology complexity in physical realizing a cell, and hence cost, will be the deciding factor in most applications.

770 THE HETEROSTRUCTURE BIPOLAR JUNCTION TRANSISTORS (HBJTs or HBTs)

The preceding sections of this chapter have shown that several fundamental material and device parameters are limiting the electrical performance of the standard homogeneous Si BJT, such as the small-signal cutoff frequencies and large-signal switching speed. Illustrated by the numerical examples shown in Table 755.1, the principal parameters that limit the circuit performance were: the emitter injection efficiency, emitter diffusion-drift transit delay, the base diffusion-drift transit delay, the charging delay of the collector-base space-charge layer capacitance through the lateral base resistance, and drift transit delay through the collector-base space-charge layer. These fundamental limitations in a standard homogeneous BJT can be mostly overcome by using **heterogeneous** structures whose emitter, base and collector layers contain different chemical elements such as Si/Ge$_x$Si$_{1-x}$/Si where x is the atomic fraction of germanium in the germanium-silicon binary crystalline layer. The layers have different energy gaps (and different intrinsic and minority carrier densities), effective masses, dielectric constants, mobilities-diffusivities, and lifetimes. These parameters can be tailored to give the highest performance bipolar junction transistors and also field-effect transistors. Such heterogeneous BJTs are known as the **heterojunction bipolar transistor** (HBT) or **heterostructure bipolar junction transistor** (HBJT). The technology of tailoring the atomic-chemical composition of thin layers to vary the energy gap has been popularly known as **energy-gap engineering** (or bandgap engineering). This term states only one important reason of the improvement - control of minority carrier concentration via energy gap. It disguises the control of the other two equally important transport properties and material parameters: mobility-diffusivity and lifetime. For integrated optics applications, the optical properties of the semiconductor (such as the complex permittivity or dielectric constant and other optical constants) can also be tailored via controlling the atomic composition of the semiconductor layers. Finally, as a result of energy gap or minority carrier density control, the majority carrier or dopant impurity concentration can also be varied which gives further control of the conductivity and transit time delay in the thin layers.

771 Historical Background

The original idea of using semiconductor layers of different energy gaps to achieve superior transistor performance was described in Shockley's 1948 transistor patent [771.1] and stated in the second claim of this patent. The proposed transistor has a emitter of larger energy gap than that of the base to give higher emitter injection efficiency. (*The grammatically correct term is 'energy gap', which is the antonym of energy band. It was used by Shockley. Band gap or bandgap are misnomers, proliferated by engineers later, who have also used the descriptives, wide and narrow bandgaps. We shall use the grammatically and technically correct descriptives **large** and **small** for **energy gap** and reserve the words wide and narrow for the geometry or areal dimensions such as the wide and narrow emitter or collector junction areas.*) Other superior and improved performance characteristics of the

HBJT's were discovered and described by Kroemer (formerly at RCA Sarnoff Research Laboratories and a professor of Electrical Engineering at the University of California at Santa Barbara since 1976) in a series of articles since 1957 [771.2]-[771.5], including the **double heterostructure BJT** in which both the emitter-base and collector-base junctions are heterojunctions. The second heterojunction (the collector-base junction) offers additional design flexibilities to attain the highest frequency in linear amplification and highest speed in large-signal digital switching. We shall describe these in the following sections.

[771.1] W.B.Shockley,"Circuit element utilizing semiconductive materials," U.S.Patent 2569347. Filed Jun.26,1948, granted Sep.25,1951, expired Sep.24,1968. (17-year statue of limitation.)

[771.2] Herbert Kroemer, "Theory of a wide-gap emitter for transistors," Proc.IRE, 45(11), 1535-1537, Nov.1957.

[771.3] Herbert Kroemer, "Heterostructure bipolar transistors and integrated circuits," Proc.IEEE, 70(1), 13-25, Jan.1982.

[771.4] Herbert Kroemer, "Heterostructure bipolar transistors: What should we build?" J.Vacuum Science & Technology, B1(2), 126-130, Apr.-Jun.,1984.

[771.5] Herbert Kroemer, "Two integral relations pertaining to the electron transport through a bipolar transistor with a nonuniform energy gap in the base region," Solid-State Electronics, 28(11), 1101-1103, Nov.1985.

[771.6] W.P.Dumke, J.M.Woodall, and V.L.Rideout (IBM), "GaAs-GaAsAl heterojunction transistor for high frequency operation," Solid-State Electronics 15(12),1339-1345, Dec.1972. (Reported the first single crystal HBT.)

The first successful single crystal HBJT was fabricated by Dumke, Woodall, and Rideout in 1972 at the IBM Research Laboratory [771.6] using the AlGaAs/GaAs heterojunction to give the larger emitter energy gap. Since then, the anticipated superior performance of HBTs has been demonstrated in GaAs-, InP- and Si-based structures whose layers are grown by the molecular-beam epitaxy (MBE) and chemical vapor deposition (CVD) techniques. Single-device HBJT examples include the lattice mismatched Si/Ge_xSi_{1-x}/Si ($f_t \propto 97$GHz); and the lattice-matched compound HBJTs such as (i) n-Al_xGa_{1-x}As/p-GaAs/n-GaAs ($f_t \propto 220$GHz), (ii) n-InP/p-InGaAs/n-InGaAs/i-InP ($f_{t\text{-exp}} = 165$GHz and >386GHz predicted) and (iii) other HBJTs using InP, AlInAs, GaInAs, and GaAlSb. Experimental small-scale-integrated digital circuits have been clocked at about 50GHz. Individual HBTs have been successfully manufactured and used in microwave and millimeter wave oscillators and amplifiers. However, digital applications have been nil due to difficulties in fabrication. Attempts to integrate the GaAlAs/GaAs HBJTs with Si-based integrated circuits have not been successful due to the large lattice mismatch, from the very different host atom species and also different lattice constants.

The status of a manufacturable monolithic HBT technology for Si based VLSI circuits has been greatly advanced in the last two years (started about 1988) by the pioneering effort of Meyerson (started about 1984) and the transistor engineers at the IBM Research Laboratory in Yorktown Heights, New York to grow thin GeSi crystalline films on Si substrates. The energy gap of GeSi is smaller than that of Si (but larger than that of Ge) so that BJTs with smaller energy gap in the base layer than in the emitter and collector layers can be made. In fact, the energy gap can be varied by varying the Ge concentration in the film to tailor the built-in electric field in the base layer. The manufacturable technology perfected by the IBM researchers is known as UHV/CVD LTE (Ultra-High-Vacuum $<10^{-9}$Torr, Low-Pressure Chemical Vapor Deposition $\sim 10^{-7}$Torr, Low Temperature Epitaxy $\sim$500C). It epitaxially grows strained thin Si/Ge_xSi_{1-x} double-layers and Ge_ySi_{1-y}/Ge_xSi_{1-x} multilayer superlattices (known as multiquantum wells, MQW) on large-area single crystal Si substrates. The subscripts x and y are the atomic percent of Ge in the film. The layers are nearly perfect (zero defect such as misfit dislocations) and highly stressed. The stress is biaxially contractile if the interatomic spacings of the lateral plane are smaller than the equilibrium biaxial spacings. The stress is biaxially tensile if the interatomic spacings in the lateral plane are larger than the equilibrium biaxial spacings. The layers can be very thin (<10A or a few atom or monolayer thick). A critical thickness of 1-μm (about ten thousand atomic or monolayers) at $x \sim \lesssim 0.1$ have been attained before defects (misfit dislocations) appear which would degrade the transistor characteristics. Grading of the Ge concentration will increase the critical thickness. The layers can be doped to high concentrations of donor or acceptor impurities during the LPCVD film growth. Both n/p/n and p/n/p double heterojunction BJTs of the Si/Ge_xSi_{1-x}/Si structure have been built to give a smaller energy gap in the base layer. The latest performance figures using this technology were reported by IBM researchers at the December-1990 International Electron Device Meeting. Performance data of non-self-aligned n/p/n Si/Ge_xSi_{1-x}/Si HBJTs included: ideal base current without the emitter-base junction space-charge layer recombination component [indicated by $I_B \propto \exp(qV_{EB}/nkT)$ with $n=1.00$ down to the lowest measurable base current], $\beta_{peak} > 100$ @300K and $\rightarrow 150$ @77K, f_t(at $h_{21}=1$) $\propto$ 75GHz(300K) and 94GHz(85K), and t_{ring} = 24.6ps at 3mW/gate from unloaded-ECL ring oscillators with active pull down and 22ps at 4mW in ECL circuits. Performance data of non-self-aligned Si/Ge_xSi_{1-x}/Si p/n/p HBJTs included: $\beta_{peak}=65$, $f_t \propto$ 30GHz(measured) and $f_t \rightarrow$ 50GHz(calculated).

The rapid recent advances in the Ge_xSi_{1-x} HBJT technology are the results of three fabrication techniques developed during the last decade (1980-1989). Highly perfect thin Ge_xSi_{1-x} films are fabricated from which the basic understanding of electronic energy bands and film growth mechanisms are developed. The basic research shows that the thin layer (with x-axis or c-axis as the normal) is **coherently** and **biaxially strained** (stretched or contracted) in the y and z directions or y-z plane), **tetragonally distorted** or **deformed** (lattice constant: $a=b\neq c$ and interaxial angle: $\alpha=\beta=\gamma=90^\circ$), **metastable** (actually stable up to a critical temperature and thickness), **single crystalline** (has no misfit dislocations and point defects), and **pseudomorphic**, all achieved from **commensurate growth**. The preceding boldfaced terms are commonly used in recent literature by researchers to describe the electronic conduction and electromechanical properties of the layer. Some of the terms are new additions to the standard ones used in textbooks of metallurgy and electronic material. The term 'commensurate' means equal in extent, i.e., the lateral lattice constants a_y and a_z of the film are exactly the same as that of the seed or substrate. This results in large strain in the plane of the film due to the stress. The film is stretched if $a_{o\text{-substrate}} > c = a_{o\text{-film}}$ and contracted if $a_{o\text{-substrate}} < c = a_{o\text{-film}}$. The subscript 'o' denotes the stress-free equilibrium value of isolated substrate and isolated film. The film thickness or interatomic spacing along the c-axis is determined by minimizing the equilibrium elastic energy. The fabrication methods and device physics are described in the next two sections, 771 and 772. In the section 773, the effects of tetragonal deformation on the energy bands and phonon spectra will be illustrated using the atomic displacement picture to show the change of the Coulomb forces between the atomic cores and on electrons.

The term 'alloy', such as in 'alloy layer', has been used by some authors to denote the crystalline Ge_xSi_{1-x} layer and other crystalline compound semiconductor layers while other authors use the term 'alloy' only for noncrystalline layers. The use of the word 'alloy' to denote a crystalline layer is a misnomer since it has been traditionally used by metallurgists and chemists at least since 1598 (when it was used in connection with coinage, as recorded by the Oxford English Dictionary) to denote a solid material piece that contains a **random mixture** of two or more elements, solid phases, molecule types or compounds, or many **randomly oriented** crystallites of different grain sizes joined by grain boundaries. Macroscopic random features of alloys were first observed via optical and electron microscopy. The misnomer 'alloy' was adopted by later engineers to describe compound semiconductor devices following R.N.Hall and W.C.Dunlap who made the first Ge homojunction diodes and transistors by the **alloying-diffusion process** in 1950 at the General Electric Research Laboratory. These Ge diodes and transistors indeed contained a thick polycrystalline emitter layer which is obtained during the cooling stage of the alloying process. However, a thin regrown layer on the substrate Ge is produced during cooling. It is crystalline and essentially perfect without defects owing to the very low cooling rates which is known later as the liquid phase epitaxy layer. The pseudomorphic commensurately grown layers obtained during the last few years are single crystalline and not polycrystalline and hence not random alloy layers. A non-pseudomorphic or noncrystalline incommensurate layer can also be produced which is polycrystalline or defective and contains checkerboard arrays of misfit dislocations. It can be appropriately termed an alloy layer according to the broad definition of the term alloy. Some authors make this distinction. In this book, we shall make this distinction and will not use the term 'alloy' indistinguishably for both a strained but perfectly crystalline thin film layer and a relaxed polycrystalline thin film layer that has many misfit dislocations and defects. Only when the stress is removed by heat treatment which generates many misfit dislocations to release the stress, will the epitaxially grown film be called an alloy layer. Such an annealed film is really polycrystalline, containing many crystallites of various grain sizes rather than an alloy of random mixture of fine (small-volume) clusters each containing two or more elements or compounds.

The rapid advancement in the performance of Ge_xSi_{1-x} HBJTs in a such a short time span, less than five years (from Meyerson's first article on growth in 1986 to IBM's five reports at the December-1990/IEDM on 100GHz GeSi HBJTs), is also the result of the basic theories and experiments which have provided the fundamental understanding of the pseudomorphic film growth mechanisms and the electronic energy bands and phonon spectra of the thin, strained, pseudomorphic, single crystalline GeSi films. The earlier (up to 1986) theoretical and experimental research (on films made by the Molecular Beam Epitaxy or MBE method) have been reviewed by Roosevelt People of ATT Bell Laboratories [700.7] which also contains his many important personal theoretical and experimental contributions on energy band alignment or lineup and growth mechanisms. The review included: (i) the energy band alignments or lineups [straddling (type I), staggered (type II) and broken (type III)] and the Si/Ge_xSi_{1-x} energy band lineups first explained by Abstreiter of University of München in 1985 [771.8] which provided the theoretical breakthrough in the understanding of the electron energy band lineup, (ii) conduction and valence band offsets, and (iii) growth mechanisms such as the stress-strain magnitudes, critical thickness of perfect film due to elastic limit, and water/oxygen induced growth defects. The number of research and engineering articles on Ge_xSi_{1-x} has exploded in the last twelve months (1990). The lastest results on energy band calculations are reported in the Physical Review, Physical Review Letters, Journal of Applied Physics, and Applied Physics Letters. These have included the conditions to give a strained film with a direct energy gap in order to give efficiency light emission. This was theorized by Froyen, Wood and Zunger of SERI (Solar Energy Research Institute of the U.S. Department of Energy), based on the observation of direct optical transition by Pearsall, Bevk, Feldman, Ourmazd, Bonar and Mannaerts of ATT Bell Laboratories in 1987. In early 1990, a very intense (31% internal quantum efficiency) photoluminescence was observed Noël, Rowell, Houghton and Perovic of the Canadian National Research Council at 820meV or 1.51μm in a MQW GeSi superlattice with 40 strained Si(109A)/$Ge_{25}Si_{75}$(73A) layers grown by MBE at 400C [771.9]. The latest articles have also reported studies on the crystal growth mechanisms (critical thickness and annealing-relaxation kinetics) of commensurately grown single layer and superlattices (multiple layers or sandwich of Ge_xSi_{1-x} and Ge_ySi_{1-y} films). The HBJT performance milestones are reported in the Technical Digest of the annual IEDM (International Electron Devices Meeting), and the Electron Device Letters.

[771.7] Roosevelt People (ATT-Bell Labs), "Physics and applications of Ge_xSi_{1-x}/Si strained layer heterostructures," IEEE J. Quantum Electronics, QE-22(9), 1696-1710, Sep.1986.

[771.8] G.Abstreiter, H. Brugger, T.Wolf, H.Jorke, and H.H.Herzog, (U.München) "Strain-induced two-dimensional electron gas in selectively doped /Ge_xSi_{1-x}Ge/Si superlattice," Physical Review Letters, 54(22), 2441-2444, 3 June 1985.

[771.9] J.-P.Noel, N.L.Rowell, D.C.Houghton, and D.D.Perovic (Canadian NRC), "Intense photoluminescence between 1.3 and 1.8μm from strained $Si_{1-x}Ge_x$ 'alloys'," Appl.Phys.Lett. 57(10), 1037-1039, 8 Oct. 1990.

772 Fabrication Methods of Ge_xSi_{1-x} HBJT

The three methods of growing pseudomorphic, strained, thin, and single-crystalline Ge_xSi_{1-x} layer and multilayer (MQW) superlattices on a single crystalline Si or Ge substrate or on a thick strain-free or thin strained Ge_ySi_{1-y} buffer layer are: (1) MBE (Molecular Beam Epitaxy) pioneered by John C. Bean at ATT Bell Laboratories and at other laboratories worldwide [772.1]; (2) LRP-CVD (Limited Reaction Process) developed by James F. Gibbons and associates at Stanford [772.2]; and (3) UHV-CVD-LTE invented and perfected by Bernard S. Meyerson at IBM-Yorktown Heights Research Laboratories [772.3].

[772.1] John C. Bean (ATT Bell Labs.), "Silicon molecular beam epitaxy: 1984-1986," J. Crystal Growth 81, 411-420, 1987. See also People's review [771.1].

[772.2] James F. Gibbons, C.M.Gronet, and K.E.Williams (Stanford), "Limited reaction processing: Silicon epitaxy," Applied Physics Lett.47(7),721-723, 1 Oct.1985; T.I.Kamins, et.al.(Hewlett-Packard Co.) and J.F.Gibbons, et.al.(Stanford), "High-frequency $Si/Si_{1-x}Ge_x$ heterojunction bipolar transistors," 1989-IEDM, 647-659.

[772.3] B.S.Meyerson(IBM), "Low-temperature silicon epitaxy by ultrahigh vacuum/chemical vapor deposition," Applied Phys. Lett.48(12), 797-799, 24 March 1986. See also his succinctly written article, "Low-temperature Si and Si:Ge ultrahigh-vacuum/chemical vapor deposition: Process fundamentals," IBM J. Research and Development 34(6), 806-810, Nov. 1990.

To develop the MBE technique of method (1), considerable effort was necessary to modify the lower-temperature ultra-high-vacuum III-V MBE equipment in order to grow the Ge_xSi_{1-x} films which require higher temperatures. MBE is not a mass production technology, however, it has been used by all the fundamental researchers at industrial and university laboratories to study the energy band properties and growth kinetics of the Ge_xSi_{1-x} film. A history of the Si MBE for 1984-1986 was given by John C. Bean [772.1]. Method (2) and especially (3) have bypassed the difficulties, complexities and low throughput inherent in the MBE method of (1). Method (3) of Meyerson appears to be the manufacturable technique that can give easy monolithic integration with million-transistor Si integrated circuit chips for ultrahigh-speed digital and ultrahigh-frequency analogy applications. It has given the fastest individual (not integrated) Si and Ge_xSi_{1-x} BJTs reported to-date with $f_t \rightarrow 100GHz$ and $\rightarrow < 10ps$ gate delay.

Detailed construction of the vacuum growth apparatus and the conditions (ambient, temperatures, and times) are described in the original articles of the three methods [772.1-772.3]. The most important common key is the atomic cleanliness of the growth surface since the starting substrate surface is the seed of crystalline film growth. Thus, it must be free of any atomic contaminants. The most frequent surface contamination before inserting the substrate Si wafer into a high vacuum chamber is the random islands of one or few monolayers of residual silicon oxide which inhibits film growth at these islands. Other surface defects and impurities may be nucleation sites or growth barriers that enhance or retard the film growth and cause defect and impurity inclusions in the film. These are usually removed in the vacuum chamber prior to film growth by Argon or Helium ion bombardment and by reaction with hydrogen, or they are removed before loading the Si wafer into the vacuum chamber, by a H_2O_2 and then H_2O/HF rinse which produces a hydrophobic, air-stable, hydrogen-passivated Si surface that is oxide-free [772.3]. To prevent oxide formation during film growth, the vacuum chamber must be free of oxygen and water vapor. The latter is the major constitute of the residual pressure in the vacuum chamber. Actually, the oxygen and water partial pressures only need to be kept below certain values at the film growth temperature so that dissociation of surface silicon oxide (by hydrogen coming from the decomposition of SiH_4, GeH_4, PH_3 and AsH_5, $SiO + 2H \rightarrow Si + H_2O\uparrow$), is faster than the oxidation of the surface silicon (by the residual oxygen, $Si + O \rightarrow SiO$). Thus, the keys to grow perfect thin epitaxial films are the maintenance of (i) a atomically clean Si substrate surface to serve as the seed and (ii) a growth ambient with low oxygen and water partial pressures. Good films in a two-layer or multi-layer (MQW) superlattice structures have been grown in-situ, i.e. without taking the wafer out of the vacuum chamber into the room ambient between the growth of each layer to avoid surface contamination. However, this prevents the fabrication of integrated circuits that require multilayers separated by other semiconductor or dielectric films and processing steps (ion implant and other doping and oxidation). Thus, the eventual manufacturing clean room in the future would probably be a continuous line completely enclosed in a high vacuum pipe of one or two meters in diameter which is compartmentalized by valved and individually evacuable pass-through vacuum buffer sections to prevent cross contamination between compartments. The whole sequence of fabrication steps (wafer insertion, movements, cleaning/etching, oxidation, diffusion, epitaxy growth, patterning, etc.) would be robotically controlled and monitored by computers.

773 Operation Principle of HBJTs

The parameters that give the superior electrical characteristics to the HBJTs with the $Si/Ge_xSi_{1-x}/Si$-substrate structure are given in the following numbered list which will be discussed more quantitatively later using the transistor equations derived in sections 73n. (i) The larger energy gap of the Si emitter and the smaller energy gap of the Ge_xSi_{1-x} base, $\Delta E_{G-EB} = E_{G-E} - E_{G-B} > 0$, give large (or near unity) emitter injection efficiency. This is due to the larger equilibrium and nonequilibrium minority carrier densities in the quasi-neutral base than in the emitter. The base/emitter minority carrier density ratio increases as the temperature is lowered which increases the CE current gain. IBM has reported $\beta = 100$ at 300K and 150 at 85K for the 75-100GHz GeSi n/p/n HBTs. (ii) The energy gap difference appears mainly as the valence band discontinuity, $\Delta E_{V-EB} \propto \Delta E_{G-EB} \propto$ 150 to 200meV while the conduction band discontinuity is rather small, $\Delta E_{C-eb} \propto$ 20meV. This helps to enhance (i). (iii) Grading the germanium concentration Ge_x in the base layer gives smooth $E_C = E_C(x)$ and $E_G = E_G(x)$ and eliminates the energy spikes, notches, or pockets in $E_C(x)$ and $E_V(x)$ whose presence would retard the diffusion and drift of the minority carriers through the base layer. (iv) Grading of Ge_x through the base layer with a retrograde Ge concentration profile also gives a **built-in electric field** which greatly decreases the base diffusion transit time. This built-in electric field is sometimes known as the quasi-electric field but it is a real built-in equilibrium electric field and not a quasi field as implied by the meaning of 'quasi' first used by Shockley in 'quasi-Fermi levels' to denote the non-equilibrium Fermi levels. Base diffusion delay of holes has been reduced substantially by this built-in electric field, giving $f_t=30$GHz in p/n/p SiGe transistors (reported by IBM at IEDM-1990). (v) Lower emitter impurity doping gives lower C_{ebt} and higher BV_{EBO}. (vi) Lower emitter impurity doping also prevents or reduces the emitter energy gap shrinkage due to heavy doping effect. This further improves the injection efficiency. (vii) Higher base impurity dopant concentration, N_{BB} in p/n/p or P_{BB} in n/p/n, prevents freeze-out or impurity deionization at 77K which maintains a low base resistance and helps to realize the higher cut-off frequency and switching speed at 77K. (viii) Higher N_{BB} or P_{BB} also allows thinner base to give higher speed, higher cutoff frequency, smaller Early effect, and higher punch-through voltage. (ix) Higher N_{BB} or P_{BB} also reduces the minority carrier current injected into the base layer from the collector when the collector-base junction is forward biased in saturated switching circuits. This reduces the minority carrier storage time in the base. (x) Graded impurity doping in the base gives an additional built-in electric field that further reduces the base transit time. (xi) Higher majority carrier mobility in Ge_xSi_{1-x} gives lower $r_{b'}$. And higher minority carrier mobility in Ge_xSi_{1-x} gives smaller base transit time, $t_B=X_B^2/2D_B$. (xii) Lower collector impurity dopant concentration, N_{CC} or P_{CC}, reduces C_{cbt} and increase BV_{CBO}. (xiii) Nearly perfect films almost free of bulk and perimeter recombination centers give very high beta at low currents. (xiv) Larger Si/Si than Si/Ge_xSi_{1-x} built-in potential barrier reduces the forward injection by the Si/Si emitter-base perimeter junction, further reducing the perimeter recombination current and increasing the low-current beta. (xv) Large Si/Si barrier also reduces the current of the underlap base-collector diode when forward biased in saturation, thus, further reducing minority carrier storage in the base.

The ideal energy band diagrams of the five-layer n/p/n and p/n/p $Si/Ge_xSi_{1-x}/Si$ double heterostructure bipolar junction transistors (DHBJTs) are shown in Fig.773.1(d). They assume spatially constant concentration of the dopant impurity and Ge in base layer shown in Figs.773.2(b) and (c). The ideal n/p/n in figure (d) assumes that the energy band offset occurs as a discontinuity only in the valence band edge at the two heterojunction interfaces, i.e., $\Delta E_{C-EB}=\chi_B-\chi_E=0$, $\Delta E_{C-BC}=\chi_C-\chi_B=0$, and $\Delta E_{G-EB}=\Delta E_{G-BC}=\Delta E_V\neq 0$; and $\Delta E_C=0$. A similar assumption is made for the ideal p/n/p in this figure with $\Delta E_V=0$ and $\Delta E_C\neq 0$.

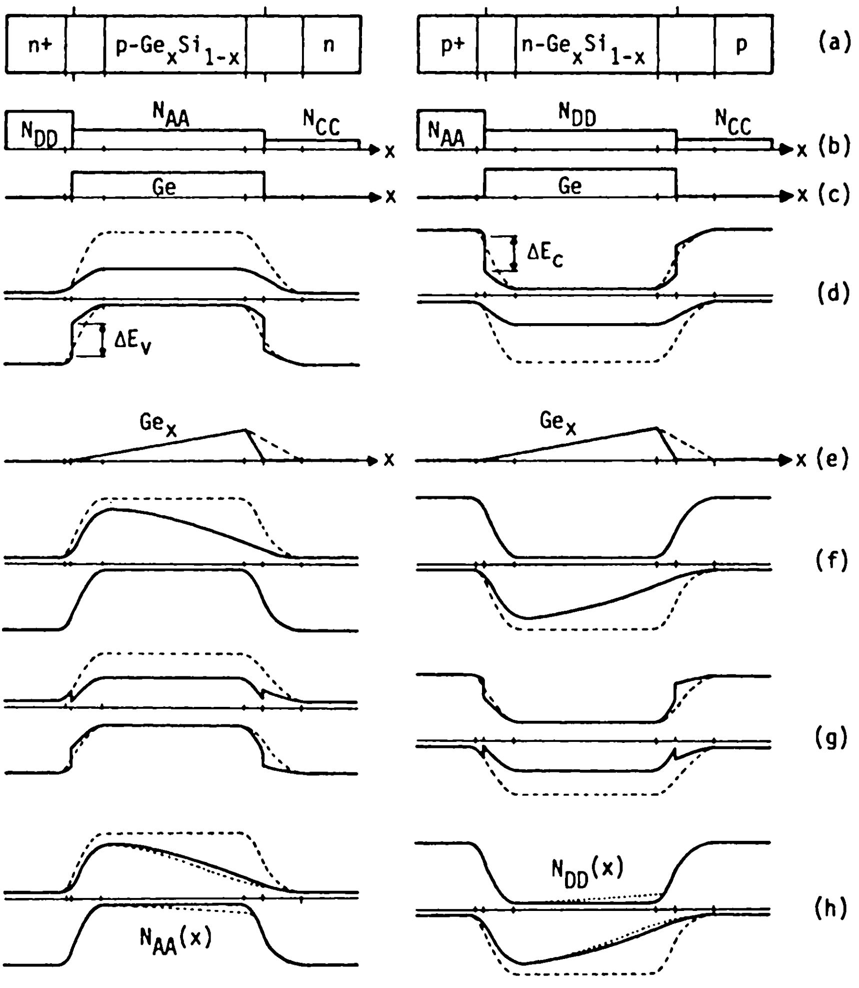

n+
p-Ge$_x$Si$_{1-x}$
n
p+
n-Ge$_x$Si$_{1-x}$
p
(a)
N_{DD}
N_{AA}
N_{CC}
x
N_{AA}
N_{DD}
N_{CC}
x
(b)
Ge
x
Ge
x
(c)
ΔE_V
ΔE_C
(d)
Ge$_x$
x
Ge$_x$
x
(e)
(f)
(g)
$N_{AA}(x)$
$N_{DD}(x)$
(h)

Fig.773.1 Energy band diagrams of Si/Ge$_x$Si$_{1-x}$/Si double heterostructure BJTs (solid lines) and homo-Si BJTs (dashed lines). (a) The physical cross-sectional views. (b) The constant doping impurity profiles. (c) The constant Ge profile in the base layer. (d) The energy band diagrams for constant-Ge$_x$ with $\Delta E_C = 0$ in n/p/n and $\Delta E_V = 0$ in p/n/p. (e) A triangular retrograde Ge$_x$ profile in the base layer. (f) Energy band diagrams with graded Ge$_x$ profile given in (e) and assuming ($\Delta E_C = 0, \Delta E_V \neq 0$) in n/p/n and ($\Delta E_C \neq 0, \Delta E_V = 0$) in p/n/p. (g) Constant-Ge$_x$ energy band diagrams with $\Delta E_C \neq 0$ and $\Delta E_V \neq 0$. (h) Retrograde-Ge$_x$ energy bands (dotted lines) with graded-$N_{BB}(x)$.

There has been considerable discussion in the literature about the theory and experiments on the energy band lineups at a heterojunction interface. The simple fundamental description of the origin of the energy bands given in sections 181 and 182 showed that the energy bands arise from the periodic potential formed by the atomic potentials due to the Coulomb force. From this, it is easy to conclude that the discontinuities are

$$\Delta E_{C-EB} = \chi_B - \chi_E \tag{773.1}$$

and

$$\Delta E_{V-EB} = (\chi_B + E_{G-B}) - (\chi_E + E_{G-E}). \tag{773.2}$$

These relationships have been used to line up the energy bands at the two heterojunctions of the Metal/Oxide/Semiconductor capacitors and transistors in chapters 4 and 6, and at the Metal/Semiconductor heterojunction of the Schottky-barrier Bethe diode in chapter 5. The fundamental **electron affinity difference** (EAD) relationships given by (773.1) and (773.2) were used first by Shockley in the late 1940's and early 1950's. They were also used by the initial researchers on Schottky barriers but not as explicitly stated and not in modern notations and terminology. (See Spenke's book.) The first extensive application of (773.1) and (773.2) was made by R.L.Anderson (while at IBM now at Syracuse University) in 1962 to semiconductor heterojunctions. It has frequently been referred to in recent literature as the Anderson or Shockley-Anderson electron-affinity model (EAM). It is not a model in the usual empirical sense. Equations (773.1) and (773.2) are the fundamentally correct relationships based on first principle.

Numerical solutions of the Schrödinger equation of a crystal cannot at present give a rigorous numerical proof of (773.1) and (773.2) because the reference potential energy in each crystalline film or bulk cannot be precisely computed. The reference potential energy is the average potential energy in the nearly-free electron approximation or the tight-binding approximation described in sections 181 and 182. This uncertainty arises because the absolute periodic potential energy cannot be constructed with precision from the approximate Gaussian-shaped atomic potential recently used by the theorists. The consequence is that the **reference potential energy** of the calculated theoretical energy bands of Si is different from that of Ge_xSi_{1-x}. Thus, the two calculated energy bands (Si and Ge_xSi_{1-x}) cannot be lined up precisely in order to predict theoretically the conduction and valence band discontinuities, ΔE_C and ΔE_V. The lineup is uncertain by the unknown difference between the two reference potential energies. The confusion among the researchers on heterojunction interfaces was further accentuated and aggravated by the significant variations in the experimental electron affinity values measured (i) on a free surface or vacuum-solid interface whose potential energy barrier height or photoelectric work function is lowered by random uncontrolled surface contaminations and strains, and (ii) on solid/solid interfaces by indirect experimental methods (such as C-V and I-V characteristics) which are subject to large experimental and interpretation errors as well as uncertainties in sample composition, interfacial layer, and geometry. In principle, interfacial and surface layers do not alter the measured electron affinity or work function difference if there is no strain. This was illustrated by the energy band diagrams of the MOSC and S/M diodes of chapters 4 and 5. However, the surface and interfacial layers are likely to produce strain or alter the interatomic spacing in the interfacial or transition layer and hence change $\chi=0-E_C$ and $E_C-E_V=E_G$ near the surface or interface from their values deep inside the bulk crystal. Since the advent of ultrahigh vacuum technology using ion pumps and oiless mechanical pumps, increasingly accurate and reliable experimental results have been obtained on atomically clean and annealed or relaxed free surfaces in ultrahigh vacuum (UHV) chambers. These new experiments are giving data approaching the true electron affinity values of the relaxed or strain-free solid/vacuum interface. To include the effects of strain or lattice deformation at the solid/solid heterointerface, which shifts E_C and E_V and hence increases or decreases the work function and electron affinity, photoelectric work function measurements of the strained-layer/vacuum interface in a UHV chamber must be made.

Let us now obtain the emitter efficiency of a n/p/n HBT using the ideal energy band diagram of Fig.773.1(d) which assumes $\Delta E_C=0$ and $\Delta E_G=\Delta E_V$. The emitter efficiency was defined by (734.9B) or (737.13) in terms of the diffusion-recombination current coefficients j_E, j_{EB}, and j_B which is

$$\gamma_E = j_E/(j_E + j_{EB} + j_B). \tag{773.3}$$

To delineate the effect of a larger emitter energy gap or smaller base energy gap on the emitter injection efficiency, the space-charge layer recombination current is neglected as a first approximation by setting $j_{EB}=0$. In most III-V and other compound semiconductor heterojunctions, j_{EB} cannot be neglected because there is a very high concentration of interfacial defects. These defects are efficient electron-hole recombination centers and they cannot be avoided or annealed out by any III-V fabrication techniques to date. However, nearly perfect commensurate and strained $Si/Ge_xSi_{1-x}/Si$ interfaces have been grown by the Meyerson-IBM technique which have given the ideal Shockley-diode base current in IBM's GeSi HBJTs, $I_B \propto \exp(qV_{EB}/nkT)$ with $n=1.00$, and a defect-recombination center density of less than 10^3 center/cm^3. Thus, $j_{EB}=0$ is a realistic and good approximation. The other two techniques, MBE and LRP, still give a very large j_{EB} component at low V_{EB} at this writing (1991).

To delineate the effect of ΔE_V, we also approximate j_B and j_E by the Gummel number of the base and emitter layers, given by (737.3) and (737.11). This approximation assumes that the layer is thin compared with the minority carrier diffusion length: $X_E << L_E=\sqrt{D_E\tau_E}$ and $X_B << L_B=\sqrt{D_B\tau_B}$. It should be emphasized that the intrinsic carrier or equilibrium minority carrier concentration is higher in the smaller-gap base than in the larger-gap emitter at the same majority carrier density or dopant impurity concentration. Their ratio is given by the energy gap difference $\Delta E_{G\text{-}BE}$. It is equal to the valence band discontinuity $\Delta E_{V\text{-}EB}$ in this ideal model that assumes no discontinuity of the conduction band, $\Delta E_{C\text{-}EB}=0$, or equal electron affinity of Si and Ge_xSi_{1-x}, $\chi_{Si}=\chi_{Ge}$. For an n/p/n HBJT where $N_{EE}=N_{DD}$ and $N_{BB}=N_{AA}$, then

$$\begin{aligned} P_E/N_B &= [n_i^2(\text{Si-Emitter})/N_{EE}]/[n_i^2(Ge_xSi_{1-x}\text{-Base})/N_{BB}] \\ &= (N_{BB}/N_{EE})\exp(-\Delta E_{G-EB}/kT) \\ &= [N_{AA}(\text{Base})/N_{DD}(\text{Emitter})]\exp(-\Delta E_V/kT) \end{aligned} \qquad (773.4)$$

The emitter injection efficiency from (773.3) is then

$$\gamma_E = 1/[(j_E/j_B) + (j_{EB}/j_B) + 1] \qquad (733.5)$$

$$\approx 1/[(j_E/j_B) + 1] \qquad (733.5A)$$

or

$$\gamma_E \approx \frac{1}{1 + (D_E/D_B)(N_{BB}X_B/N_{EE}X_E)\exp(-\Delta E_V/kT)} . \qquad (773.6)$$

For spatially varying concentrations of $Ge_x(x)$, emitter dopant impurity $N_{EE}(x)$, and base dopant impurities $N_{BB}(x)$, the ratio j_E/j_B in (773.5A) is a ratio of two integrals given by (737.4) and (737.12). Then,

$$\gamma_E \approx \frac{1}{1 + \dfrac{\displaystyle\int_{0_B}^{X_B} [N_{BB}(x)/n_i^2(x)D_B(x)]dx}{\displaystyle\int_{0_E}^{X_E} [N_{EE}(x)/n_i^2(x)D_E(x)]dx}} . \qquad (773.7)$$

In the above, $N_{EE}=N_{DD}$ and $N_{BB}=N_{AA}$ for the n/p/n transistor and $n_i^2(x)$ varies with $Ge_x(x)$ in the base. This can be further generalized to give the formula of forward alpha, α_F, (and reverse alpha α_R) by including the base transport factor, α_B. α_B has the simple formula $\mathrm{sech}(X_B/\sqrt{D_B\tau_B})$ for constant base concentrations but it is altered by the two built-in electric fields [from $E_G(x)$ and $N_{BB}(x)$], and by the x-dependent mobility and diffusivity due to dopant impurity ion scattering from x-dependence of $N_{ION}=N_{BB}=N_{BB}(x)$.

The result of the constant Ge_x and N_{BB} n/p/n HBJT, given by (773.6), shows that the valence band offset, ΔE_V, due to larger emitter energy gap than base, can easily overwhelm the dopant concentration ratio, $N_{BB}/N_{EE}=N_{AA}/N_{DD}$, because of the exponential dependence, $\exp(-\Delta E_V/kT)$. This allows higher dopant (acceptor) concentration in the n/p/n base which would reduce (i) the lateral base resistance, $r_{b'}$, and (ii) the base majority carrier (holes) freeze-out at low temperatures (section

252). Freeze-out would have increased the base resistance tremendously and prevented the HBJT from operating at 77K for higher speed due to higher mobility at 77K. This ΔE_V also allows a lower dopant (donor) concentration in the n/p/n emitter which would (iii) reduce the emitter-base transition layer capacitance, C_{ebt}, (iv) lessen the reduction of the emitter-base breakdown voltage, (v) lower the emitter-base leakage current due to band-to-band tunneling because of higher base impurity concentration, and (vi) lessen the emitter Si energy gap shrinkage due to heavy doping which would decrease the injection efficiency. Since the emitter efficiency is proportional to the total dopant concentration or majority carrier concentration in the base layer, $P_{BB}X_B = N_{AA}X_B$, the higher base impurity concentration also (vii) allows thinner base layer before punch-through or considerable Early effect sets in. (viii) The thinner base also gives higher base transport factor, $\alpha_B = 1 - X_B^2/2D_B\tau_B$ and hence a larger d.c. beta. (ix) It also reduces the base diffusion transit time delay, $t_b = X_B^2/2D_B$, and further increases the cutoff frequency and the switching speed. (x) A higher minority carrier (electron) mobility in the base is also anticipated, even when the ion-scattering-limited mobility will dominate due to the high base impurity concentration, because the phonon-scattering-limited electron mobility in Ge is three times higher than that in Si.

In view of the many geometry and material parameters that can be varied due to the smaller energy-gap in the base layer of a HBJT, an optimum set of the parameter values can be found in a detailed numerical design to maximize the performance. The optimum HBJT design for each application will have different values of the geometry and material parameters. For example, in small-signal amplifications, f_t is maximized by reducing the minority carrier delay times in the five layers. On the other hand, in maximizing the maximum-frequency-of-oscillation for signal generator applications, both f_t and $1/r_b \cdot C_c$ need to be maximized. Similarly, different sets of optimum parameter values are obtained for HBJTs designed to have the highest switching speed (gate delay) and the least switching power or energy (power-delay product). The optimum switching HBJT design also depends on the circuit used, ECL, TTL or NTL.

Let us next consider the effect of position dependent energy gap arising from grading the Ge concentration. A graded Ge_x concentration profile in the base layer would create a built-in electric field that can accelerate or decelerate the minority carriers. The triangular **retrograde** $Ge_x(x)$ profile shown in Fig.773.1(e) would create an aiding built-in electric field that accelerates the minority carriers (electrons in n/p/n) from the emitter to the base. This is illustrated by the energy band diagrams in Fig.773.1(f). **Retrograde** means the grading has a positive slope near the front surface and negative slope near the back surface, like the triangular Ge profile shown in Fig.773.1(e). From the energy band diagrams of Fig.773.1(f), it is evident that the built-in field comes from the spatial dependence of $E_G(x)$ in the Ge_xSi_{1-x} base due to the varying atomic-fraction, x, of Ge_x and not from the spatially constant $E_V(x)$ in n/p/n, or $E_C(x)$ in p/n/p.

In general, the currents from the built-in electric fields due to a position dependent energy gap $E_G(x)$ and electron affinity $\chi(x)=E_C(x)$ appear in the two continuity equations of the Shockley equations. They give a built-in acceleration or deceleration electric field for electrons and a different acceleration or deceleration electric field for holes. The built-in fields are given by

$$qE_e(x) = dE_C(x)/dx = d\chi(x)/dx \tag{773.8A}$$

and

$$qE_h(x) = dE_V(x)/dx = d[E_G(x)+\chi(x)]/dx. \tag{773.8B}$$

The energy band diagram of the triangular retrograde Ge_x profile in Fig.773.1(f) shows that the built-in electric field accelerates the diffusion of minority carriers (electrons in n/p/n) through the base layer from the emitter to the collector. This reduces the minority carrier base transit time. It has little effect on the diffusion and drift of majority carriers (holes in n/p/n) through the base layer since the majority carrier concentration is spatially constant, $P=N_{AA}\neq f(x)$ in n/p/n, under low injection level conditions. High injection level diminishes the built-in field from graded $Ge_x(x)$ just like the reduction of the built-in field from graded $N_{BB}(x)$.

As an estimate, a Ge grading in the n/p/n HBJT that gives 0.15eV potential energy drop in the p-type Ge_xSi_{1-x} base layer would increase the minority carrier (electrons) diffusion current and the apparent diffusivity by $0.15/0.025=6$.

The realistic HBJT energy bands with constant Ge and dopant concentrations are shown in Fig.773.1(g) in which $\Delta E_C\neq 0$ in n/p/n and $\Delta E_V\neq 0$ in p/n/p. If we assume that the published electron affinities of Si and Ge ($\chi_{Si}=4.013$eV, $\chi_{Ge}=4.13$eV) are accurate strain-free values, then $\Delta E_C(\text{Si/Ge}) = \chi_{Ge}-\chi_{Si} = 4.13-4.013 = 0.12$eV for a strain-free Si/Ge heterojunction. For commensurately strained Si/Ge_xSi_{1-x} thin layers, ΔE_C has been found to be small, ~0.020eV, but it can still be a very significant discontinuity at low temperatures. At 77K, kT=6.63meV, and $\exp(-\Delta E_C/kT) = \exp(-20/6.63) = \exp(-3) = 0.05$. Thus, the carrier concentration and current is reduced by 20 times across the discontinuity.

The n/p/n in Fig.773.1(g) illustrates a large valence band edge discontinuity and a small but significant notch-spike-pocket discontinuity in the conduction band edge at the EB and BC junctions. The E_C-spike near the EB junction reduces the injection of the electrons (minority carriers) into the base from the emitter. The same E_C-notch-spike-pocket near the CB junction prevents the minority carrier (electrons) from being collected. The reduction effect from these spikes-pockets-notches increases exponentially as the temperature decreases, $\exp(-\Delta E_{C,V}/kT)$, when the energy band discontinuities, ΔE_C and ΔE_V, become comparable to or several-times larger than kT as the preceding numerical example just indicated. The spikes can be reduced and removed by retrograding the Ge profile and by grading the dopant impurities in the base and the collector which will give a result like that shown in Fig.773.1(f).

The effect of the spikes is much more severe for holes in the p/n/p HBJT with commensurately strained Si/Ge_xSi_{1-x}/Si structure because $\Delta E_V \simeq 0.15$-0.2eV is much larger than $\Delta E_C \simeq 0.02$eV. This is illustrated by the p/n/p energy band diagram in Fig.773.1(g) which has a constant Ge profile in the base. A retrograde Ge profile in the base can be designed to eliminate the two spike-pockets with the consequence that a large built-in aiding drift electric field for holes is also obtained. The energy band diagram of such a p/n/p HBJT is shown in Fig.773.1(f). Such an aiding drift electric field is precisely what is needed to reduce the minority carrier (hole) diffusion delay in the base layer in p/n/p Si/Ge HBJTs in order to overcome the inherently small hole mobility and to give high speed and high frequency performances. This has been demonstrated in p/n/p Si/Ge_xSi_{1-x}/Si HBJTs reported by IBM. It showed that proper Ge retrograde grading plus proper grading of the donor dopant impurity profile in the base can diminish ΔE_V as well as improve the observed experimental f_t=30GHz to 50GHz in GeSi p/n/p HBT. This offers the prospect of producing ultrafast CBiCMOS.

In addition to the new flexibility of grading the Ge concentration in the base to give an aiding built-in field or **spatial dependence of the minority carrier band edge**, $E_V(x)$ for p/n/p in Fig.773.1(f), we still have the flexibility of grading the majority-carrier dopant impurity concentration in the base which gives a second aiding built-in electric field to hasten the minority carrier diffusion in the base. The field is aiding for minority carrier diffusion through the base from the emitter to the collector if N_{DD} is larger near the EB junction than in the BC junction, i.e., a non-retrograded N_{DD} profile. This second built-in electric field appears as the **spatial dependence of the majority carrier band edge**. For p/n/p, this is the dotted $E_C(x)$ curve in Fig.773.1(h). The $N_{DD}(x)$ grading also affects $E_V(x)$ which is also shown by a dotted curve in Fig.773.1(h). A similar second built-in field can be obtained in n/p/n GeSi HBJTs by grading the acceptor concentration in the base, $N_{BB}=N_{AA}(x)$. The spatial variation of $E_V(x)$ graded-base-impurity n/p/n is shown by the lower dotted curve in Fig.773.1(h). Similarly, due to $N_{BB}=N_{AA}(x)$. $E_C(x)$ is also shifted as indicated by the upper dotted curve of the n/p/n energy band diagram in Fig.773.1(h).

774 Energy Bands and Phonon Spectra of Commensurate Layers

The effects of stress-strain on the energy bands and phonon spectra of a thin epitaxial commensurately grown layer on a semi-infinite substrate can be understood via the simple Coulomb electrostatic law without quantum mechanical calculations. We shall present this description, not available in any literature thus far, to aid in the understanding of the results of the detailed theoretical calculations and experimental results given in the recent research and engineering articles on the

Ge_xSi_{1-x} commensurate films. It is based on the simple physics presented in chapter 1 on energy bands and chapter 2 on phonon spectra.

We start with the description of the lattice distortion in a commensurate growth in order to define the terms and provide the crystal model graphically. Figures 774.1(a) (b) and (c) show the commensurate growth of a thin Ge film on a thick Si substrate while the three primed figures (a'), (b') and (c') show the commensurate growth of a thin Si film on a thick Ge substrate. Figures 774(d) and (d') give the lattice geometry in an incommensurate growth or a commensurate grown film that was annealed to relieve all the stress and strain.

For better perspective, only the corner Si (circles) and Ge (dots) atomic cores of the cubic and tetragonal unit cells are shown in Fig.774.1(a), and partially shown in Figs.774.1(b) and (c). Si and Ge atomic cores at the face-centered positions and in the interior of the unit cells are not shown. For Ge_xSi_{1-x} films with $x \neq 1$, the inter-Ge distance in the plane of the superlattice layer is larger than the equilibrium value, causing x to be discrete instead of continuous. This poses a limit on the minimum area of the film below which quantization and bound states in a 3-d well (quantum dot) would appear.

The key for a simple understanding of the physics lies in the difference of the **equilibrium** lattice constant: at 300K, $a_{Si}=5.43095A < a_{Ge}=5.64613A$, i.e. the equilibrium atomic spacing in Ge crystal is larger than that in Si crystal. To illustrate using this key, let us denote the plane of the film by the coordinates y and z. They are **transverse** to the film **growth** or **longitudinal** direction, x, which is also the direction of minority carrier transport in a HBJT.

Thus, the commensurate Ge film on Si shown in (b) is **contracted** in the y and z directions of the film plane by a **biaxial contractile stress** (force), and **elongated** or **extended** in the x direction normal to the film plane by a **uniaxial tensile stress**, making the Ge film thicker and its area smaller. The three-dimensional forces can be decomposed into a **hydrostatic pressure (triaxial compression)** force and a **uniaxial tension** force in the x-direction. This decomposition is used in the analysis of the shifts of the energy band and phonon spectra as a function of stress.

The commensurate Si film on Ge shown in (b') has just the opposite geometrical distortion: elongated or extended in the plane or y and z directions by a **biaxial tensile stress** and contracted in the x direction by a **uniaxial compressive** or **contractile stress**, making the Ge film thinner and its area larger. The three-dimensional force can again be decomposed into an **antihydrostatic** (or expansive) **pressure (triaxial tension)** and a **uniaxial contraction** in the x-direction.

Both of the preceding film examples have undergone **tetragonal distortion** or **tetragonal deformation** from its original **cubic** shape to **tetragonal** shape.

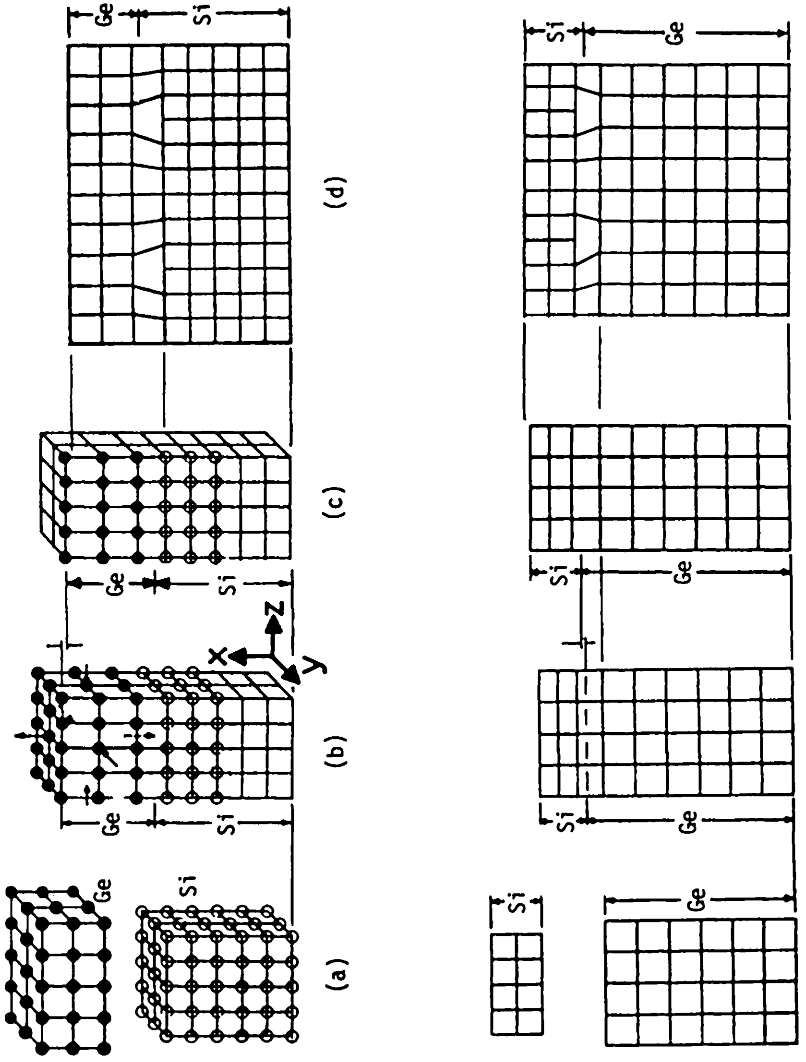


Fig.774.1 Exaggerated for clarity, views of commensurate growth of strained Ge-film on unstrained Si (unprimed), and of strained Si-film on unstrained Ge (primed). (a) and (a') Equilibrium arrays of cubic unit cells. (b) and (b') Biaxially strained film commensurately grown. (c) and (c') Graded heterointerfacial layers. (d) and (d') Relaxed/annealed incommensurate film showing two of the many misfit dislocation half-planes. [(a) and (b) from J.C.Bean.]

As indicated in Figs.774.1(c) and (c'), relaxation of the x-spacing of the atomic planes must occur over a few atomic planes in the heterointerfacial transition layer because of the geometrical asymmetry of a thin film on a semi-infinite substrate. The lateral (transverse) distortion is zero in commensurate films because large film area contains many atoms. The variation of the atomic plane spacing will determine the energy band structure and the phonon spectra at the heterointerface. Variation of the interplane spacing has not been reported in the literature.

Figures 774.1(d) and (d') show a relaxed incommensurate grown film in which the stress is released by the presence of misfit dislocation half-planes. Such a film is obtained when a defect- and dislocation-free commensurately grown strained film is annealed at a high temperature to release the stress and relax the film to its equilibrium size. It is evident that such films are electronically inferior for the following reasons. (i) The dangling bonds on the dislocations are electron and hole traps which would reduce the minority carrier lifetimes. (ii) The randomly located lateral dislocations, whether charged or neutral, would reduce the mobilities and the diffusivities of the majority and minority carriers. (iii) The dislocations are fast diffusion pipes for dopant impurities which would cause emitter-collector short-circuit in the transistor using the film as the base layer. Thus, the key to a high electronic quality film is to avoid high temperature annealing of the nearly perfect commensurately grown film in order to retard or prevent the generation or nucleation and propagation of dislocations. This obviously reduces the flexibility to fabricate complex integrated circuits containing a commensurate film. However, it is also obvious that a thin commensurate crystalline film sandwiched between two thick and strain-free crystalline films should be more stable. Such a structure has geometrical and force symmetry and hence is not likely to relax via interfacial generation of dislocation or propagation of dislocations from the perimeters caused by asymmetrical forces.

The data of **strain** in Ge_xSi_{1-x} films on the (001)Si surface have been determined experimentally by Rutherford backscattering, X-ray diffraction, and resonant Raman scattering. These are shown in Fig.774.2(a) on the following page. When the film is thin, the strain follows essentially the elastic (dislocation-free) theory, i.e. strain ~ x_{Ge}, because the strain is proportional to the change of the lattice constant from the unstressed film. The elastic limit is exceeded when the film exceeds a critical thickness due to the excessive stress which ruptures the bonds and generates misfit dislocations. The dislocations release the stress and reduce the strain. The **critical thickness**, h_c, versus x_{Ge} in single-layer Ge_xSi_{1-x} films grown on the (001)Si surface is shown in Fig.774.2(b) on the following page. D.C.Houghton and the Canadian NRC group have recently shown that **strain relaxation**, from People's curve to the equilibrium Matthews-Blakeslee curve in Fig.774.2(b), proceeds via generation/nucleation and glide propagation of the 60° a/2 [011] misfit dislocation with activation energy of 2.5eV and 2.25eV respectively. Hull and Bean of ATT Bell-Labs reported velocities of 1nm/s at 540C and 10μm/s at 870C. These are useful to estimate transistor aging rate due to change of the energy gap and discontinuity in the base layer.

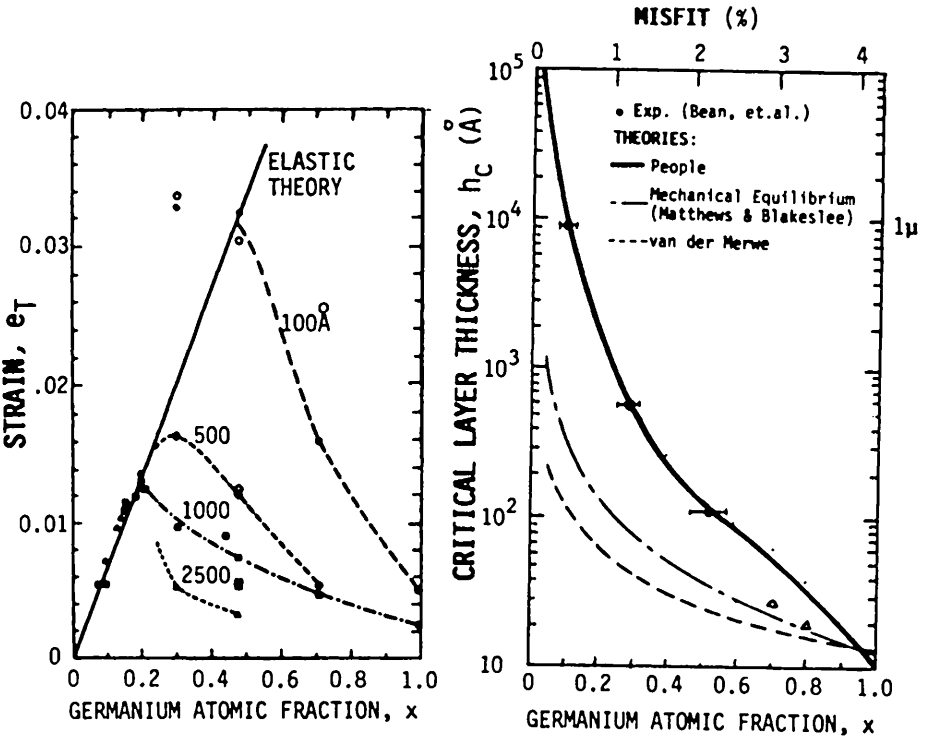
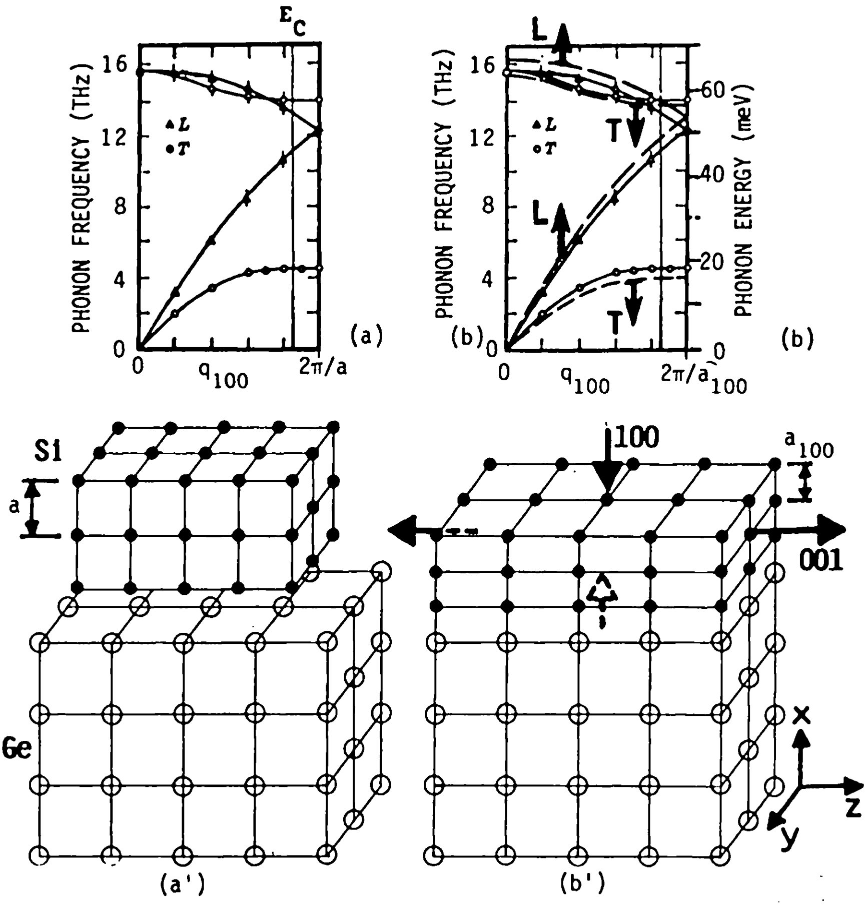


Fig.774.2 Mechanical properties of Ge_xSi_{1-x} films grown on (100)Si surface as a function of Ge fraction, x_{Ge}. (a) Strain ε_T. (b) Critical thickness, h_c. (Data from Bean et.al. and People [771.7].)

The phonon spectra of the strained commensurate films are shifted due to the lattice displacement. The shift of the optical phonon energy or frequency at q=0 (Raman frequency) has been used to measure the strain. The phonon frequency shifts can be understood from the crystal model in Fig.774.3(b'). It shows that the lattice constants in the plane of the commensurate film, $a_y=a_z$, are increased from $a_{Si}=5.431A$ to $a_{Ge}=5.646A$ while the lattice constant normal to the film or along the c-axis of the tetragonal unit cell is reduced. Larger interatomic spacing gives smaller interatomic force along the y-z directions in the film plane. Smaller force decreases the frequency of vibration of the Si atoms in the film. Thus, the ω-q curves of the two transverse modes (T-modes) in each of the two branches (acoustical and optical) are shifted downwards as indicated in Fig.774.3(a'). Similarly, the smaller interatomic distance along the c-axis or the longitudinal direction increases the interatomic force and the frequency of lattice vibration. Thus, the longitudinal mode (L-mode) of the optical and acoustical branches is shifted to higher frequencies. This explains the downward shift of the Raman frequency, $\omega_{l0}(q=0)$, observed in Raman scattering experiments during thermal annealing to reduce the strain in a commensurate Si film grown on Ge or Ge_xSi_{1-x}. This simple explanation also accounts for the upward shift of the Raman frequency of commensurate Ge or Ge_xSi_{1-x} films grown on the (100)Si surface.


E_C
PHONON FREQUENCY (THz)
16
12
8
4
0
▲L
●T
0
q100
2π/a
(a)
L
T
L
T
PHONON FREQUENCY (THz)
16
12
8
4
(b) 0
0
q100
2π/a100
60
40
20
0
PHONON ENERGY (meV)
(b)
Si
a
Ge
(a')
100
a100
001
X
Z
Y
(b')

Figs.774.3 Shift of phonon dispersion curves, ω_q-q, due to strain in a commensurate Si film grown on the (100) surface of a [100]-directed Ge substrate due to strain. (a) Unshifted in stress-free Si bulk or film. (b) Shifted in commensurate Si film. (a') and (b') are the corresponding cubic cell diagrams. (b') shows the uniaxial contractile stress (force) in the [100]- or x-direction perpendicular to the film, and the biaxial tensile stress (force) in the [010]- and [001]- or y- and z-directions in the plane of the film.

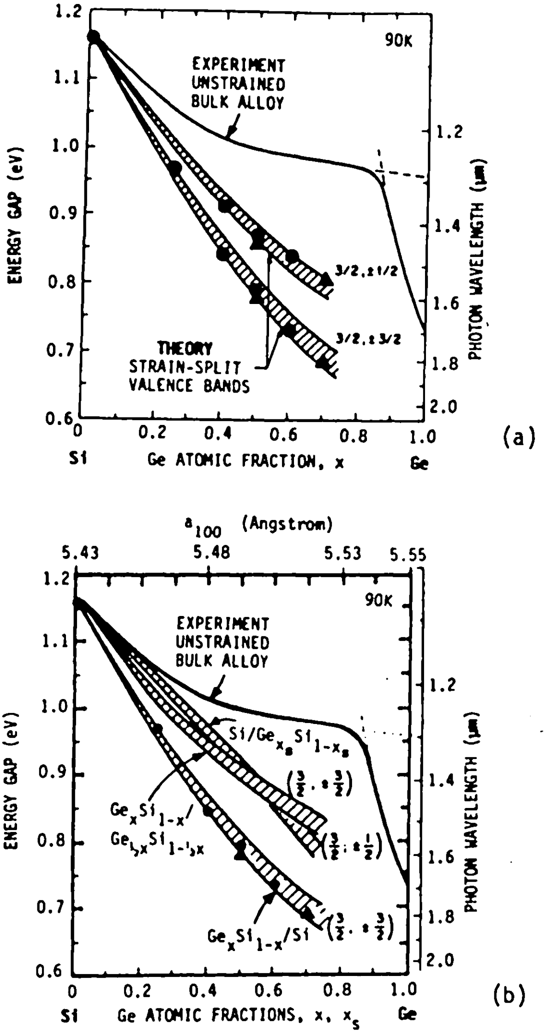


Fig.774.4 The energy gap of coherently strained commensurate GeSi films on <100> substrates. (a) The unstrained bulk alloy and the strain-split valence bands. (b) Three different substrates. (Data from Lang, People and Bean.)

The **energy bands** or the conduction and valence band edges are also shifted by the presence of stress which strains and deforms the lattice. The shifts of energy levels are expected from the diagram of energy levels versus interatomic spacing of Fig.173.1 by varying the interatomic spacing. In the elastic limit, the energy-level-shift versus strain can be derived theoretically using the deformation potential constants which also determines the phonon-scattering-limited mobilities of electrons and holes discussed in section 313 and given by (313.7) and (313.9). The experimental data and the theory are plotted in Figs.774.4(a) and (b).

The upper curve in Fig.774.4(a) is the energy gap of unstrained bulk Ge_xSi_{1-x} alloy from optical absorption measurements made by Braunstein, etal. in 1958 at RCA Laboratories when the transistor market was dominated by Ge transistors. This early date indicates that efforts were made to evaluate the possibility of using semiconductors with energy gap larger than Ge for device applications more than thirty years ago before the development of Si transistor technology. The two lower shaded energy curves are the recent experimental energy gaps between the conduction band edge and the two valence band edges optically measured on strained commensurate Ge_xSi_{1-x} films on (100)Si surface. (For a review, see the strain-free bulk-Si E-k diagram in Fig.183.1 for the light hole and a heavy hole bands that give the two valence band edges.)

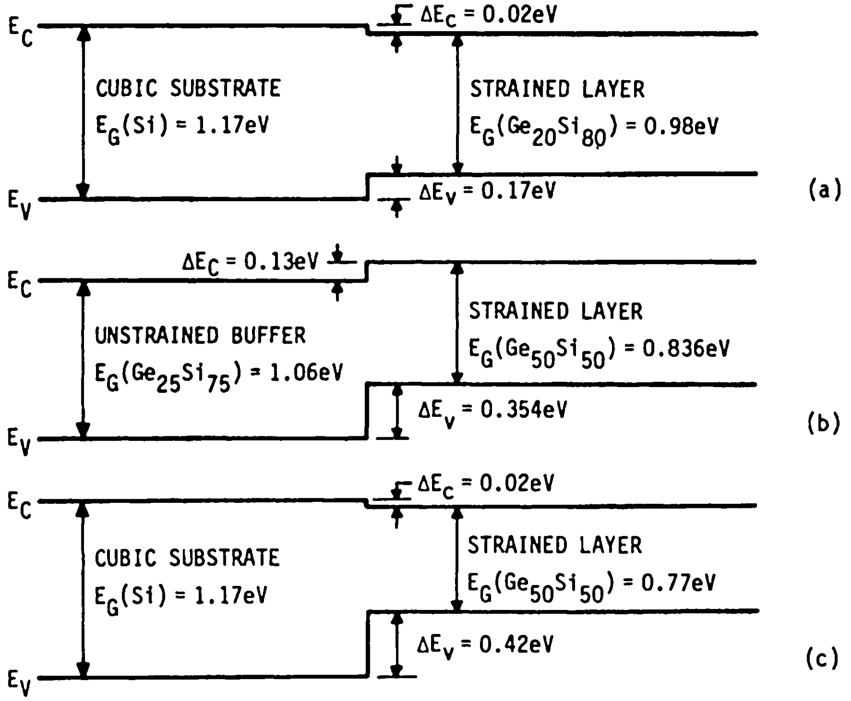
Figure 774.4(b) gives the experimental and theoretical curves of the energy gap of coherently strained (i) Ge_xSi_{1-x} film on [001]Si substrate (lower shaded curve), (ii) Si film on [001]Ge_xSi_{1-x} substrate (highest shaded curve), and (iii) Ge_xSi_{1-x} film on [001]$Ge_{x/2}Si_{1-x/2}$ substrate.

The calculated valence-band offsets for Ge_xSi_{1-x} film commensurately grown on a $Ge_{X_s}Si_{1-X_s}$ substrate can be accurately approximated by

$$\Delta E_V = (0.84 - 0.53X_s)x \qquad \text{(eV)} \qquad (774.1)$$

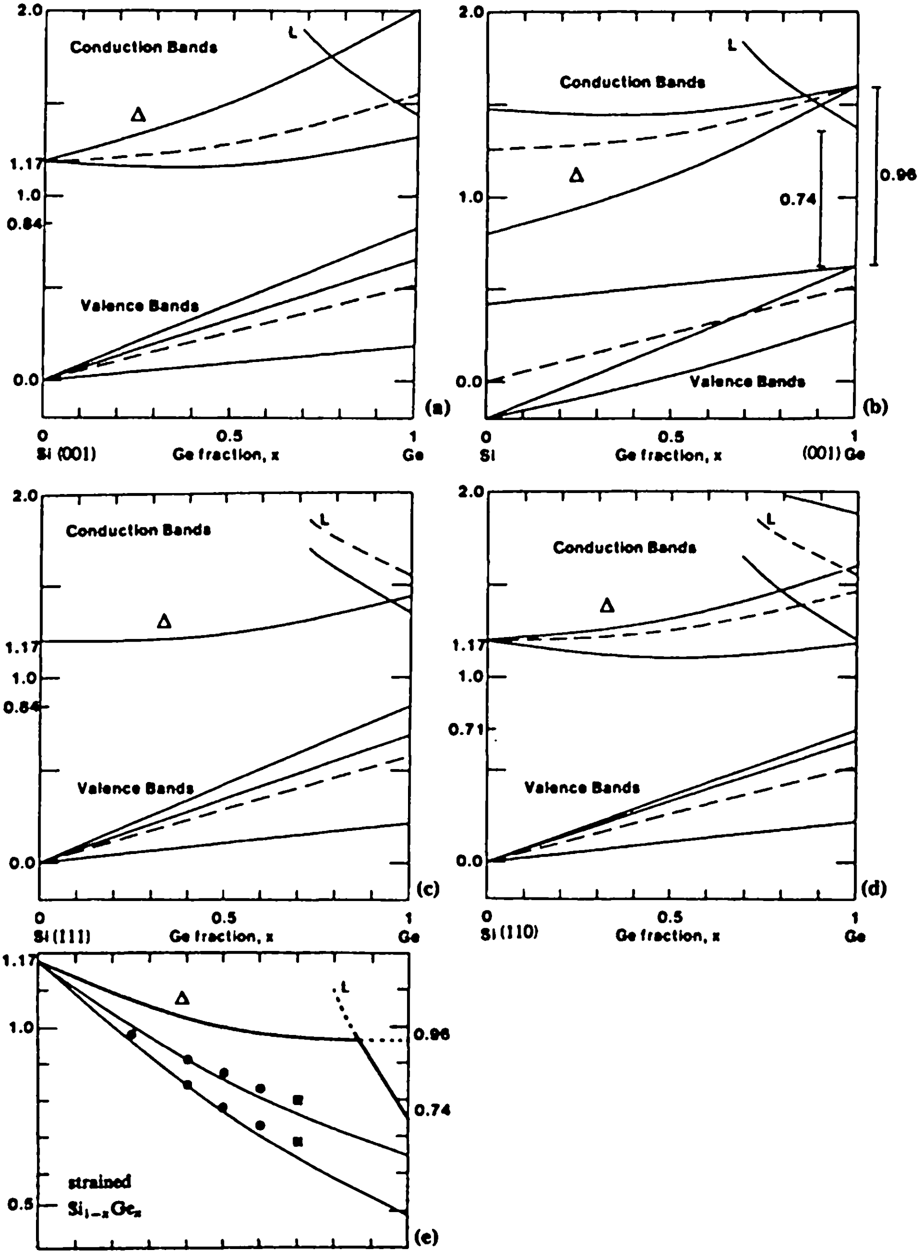
where X_s the Ge fraction in the substrate and x is the Ge fraction in the film. The value of 0.84 was a 1990 update from the 0.74 value given by People and Bean in 1986. It was based on a theoretical calculation made by Van der Walle and Martin of a strained commensurate Ge film on a [100]Si substrate.

From the energy gap data of Figs.774.4(a) and (b) and the valence band offset equation (774.1), three examples of energy band alignment were derived by People. These are recalculated using the new value of 0.84 given by (774.1) for ΔE_V and ΔE_C=0.02eV, and shown in Figs.774.5(a) to (c). Figure (a) is for a strained $Ge_{20}Si_{80}$ film on the (100) surface of cubic (crystalline) Si, known as the **type I** or **straddling** energy band alignment. Figure (b) is for a strained $Ge_{50}Si_{50}$/Si film on an unstrained [001]$Ge_{25}Si_{75}$ buffer layer, known as the **type II** or **staggered** energy band alignment. Figure (c) is for a strained $Ge_{50}Si_{50}$/Si film on [001]Si substrate, which is also a type I straddling energy band alignment.


E_C
$\Delta E_C = 0.02eV$
CUBIC SUBSTRATE
$E_G(Si) = 1.17eV$
STRAINED LAYER
$E_G(Ge_{20}Si_{80}) = 0.98eV$
E_V
$\Delta E_V = 0.17eV$
(a)
E_C
$\Delta E_C = 0.13eV$
UNSTRAINED BUFFER
$E_G(Ge_{25}Si_{75}) = 1.06eV$
STRAINED LAYER
$E_G(Ge_{50}Si_{50}) = 0.836eV$
E_V
$\Delta E_V = 0.354eV$
(b)
E_C
$\Delta E_C = 0.02eV$
CUBIC SUBSTRATE
$E_G(Si) = 1.17eV$
STRAINED LAYER
$E_G(Ge_{50}Si_{50}) = 0.77eV$
E_V
$\Delta E_V = 0.42eV$
(c)

Figs.744.5 Energy band alignments for (a) $Ge_{20}Si_{80}/Si$ heterostructure on [001]Si substrate (Type I straddling), (b) $Ge_{50}Si_{50}/si$ heterostructure on an unstrained [001]$Ge_{25}Si_{75}$ buffer layer (Type II staggering), and (c) $Ge_{50}Si_{50}/Si$ heterostructure on [001]Si substrate (Type I straddling). [After Roosevelt People, with new ΔE_V given by (774.1) and typos corrected.]

The theoretical shifts of the conduction and valence band edges as a function of the Ge fraction, x, in strained commensurate Ge_xSi_{1-x} films on Si substrate are shown in Figs.744.6(a)-(d) using the experimental data of Fig.744.6(e). These were computed by Van de Walle and Martin. The figures show that the three valence bands are split and the light hole band moves up further than the heavy hole band and the split-off band. (See Fig.183.1 for the bulk-Si E-k diagram as a reference.) The conduction band is also split and the four states in the plane of the film are shifted downwards because of biaxial contractile strain which moves the atoms closer together laterally. The smaller a_y and a_z lowers the average electron potential energy which pulls downwards the electron energy levels related to the electron transverse wave vector in the plane of the film, k_y and k_z. The larger interplane spacing along the c-axis means that the atoms are further apart in the longitudinal or c-axis direction. This increases the average potential energy which pulls upwards the energy levels related to the longitudinal wave vector, k_z. A similar basis can be used to describe the results of the other combinations shown in figures (b)-(d). The dashed lines are the average energies of all the band edges which were used by Van de Walle and Martin to empirically line up the energy band of a commensurate film with that of its substrate.



Figs.744.6 Shift of energy band edges as a function of Ge fraction in commensurate Ge_xSi_{1-x} films grown on (a) (001)Si surface, (b) (001)Ge surface, (c) (111)Si surface, (d) (110)Si surface, and (e) Experimental data. (From van de Walle and Martin, Phys.Rev.34, 5621, 1986.).

780 THE FOUR-LAYER PNPN DEVICES

Negative resistance and bistability in two-terminal vacuum and solid-state diodes have fascinated academic researchers and application engineers for decades because of their circuit simplicity (only two terminals) and application potential as: negative-resistance microwave amplifiers and oscillators, d.c. voltage regulators, two-state binary memory cells, cross-point switches in telephone switching office, light emitters with memory, and others. However, manufacturing difficulties have limited their applications as a negative resistance diode in signal applications and as two-state memory or cross-point switch. Only the gas-filled glow-discharge (or gas discharge) cold-cathode diode tube has found many applications. It was widely used until about 1960 as a d.c. voltage regulator in vacuum-tube circuits because the voltage across the tube is essentially independent of current in its high-current/low-voltage state. However, it has been continually used since 1904 as the light-emitter for illumination, known initially as the neon tube when neon filled tube was used in 1910 to produce inexpensive commercial neon signs, and later known as the fluorescent lighting tube or fluorescent lamp after fluorescent materials were used to coat the inner surface of the tube to produce the desired color. It is increasingly used in display: first as a rapidly changeable numeric display patented by H.J. Hampel in 1954 (the NIXIE tube announced by Burroughs Corp. in 1960); and now, as an array of $10^3 \times 10^3$ miniature cells in flat-panel alpha-numeric plasma display panel for computers and other display applications. About 0.1% argon has been added to neon to lower the switching or firing voltage to less than about 140V. The first plasma panel had a 50x50 array of gas-discharge plasma cells in a display area of 2'x2.5'. It was built by the Bell Telephone Laboratories in 1927 to demonstrate full gray-scale live television transmission at 16 frames per second between Washington D.C. and New York City. In 1966, Donald L. Bitzer and Hiram G. Slottow (University of Illinois) reported a major breakthrough for display applications [AFIPS Conf. Proc.29,541(1966)]. They replaced the bulky and costly discrete external high load resistor that holds the plasma cell to the d.c. high-current/low-voltage light-emitting state by an integrated dielectric capacitance under an a.c. (60Hz or higher audio) voltage drive. This not only eliminated power loss in the load resistance but can still provide memory. The dielectric capacitance is easily fabricated as an integral part of and in series with the two electrodes of each plasma cell, making the plasma display panel a commercial reality.

Multi-junction solid-state diodes have been invented as negative resistance microwave oscillators and amplifiers, such as the transit-time negative resistance Read diode and other IMPATT diodes (Chapter 10 of Sze [799.13]). The d.c. negative resistance Esaki tunnel diode (sections 154 and 570) was used as a microwave oscillator because of its extremely small time constant. The production volume of these diodes has been rather low due to the very limited application demand and the very low number of devices used in each system in addition to structural complexity and very tight material parameter requirement.

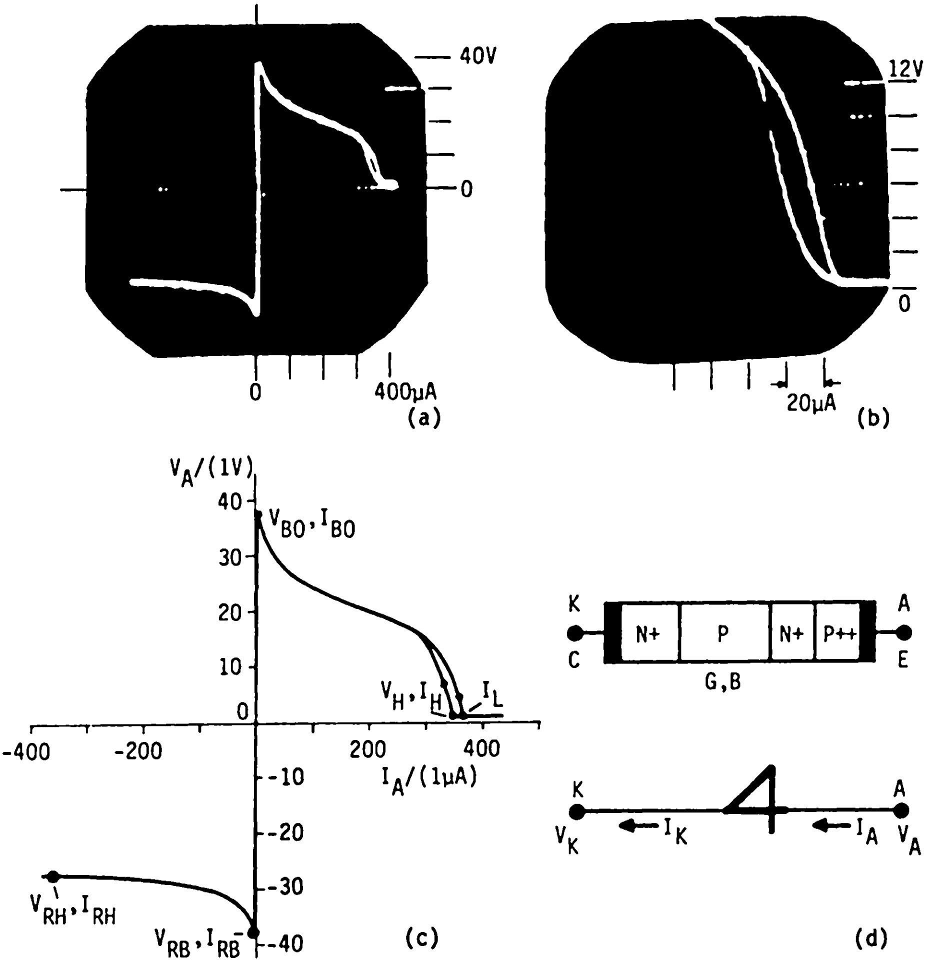
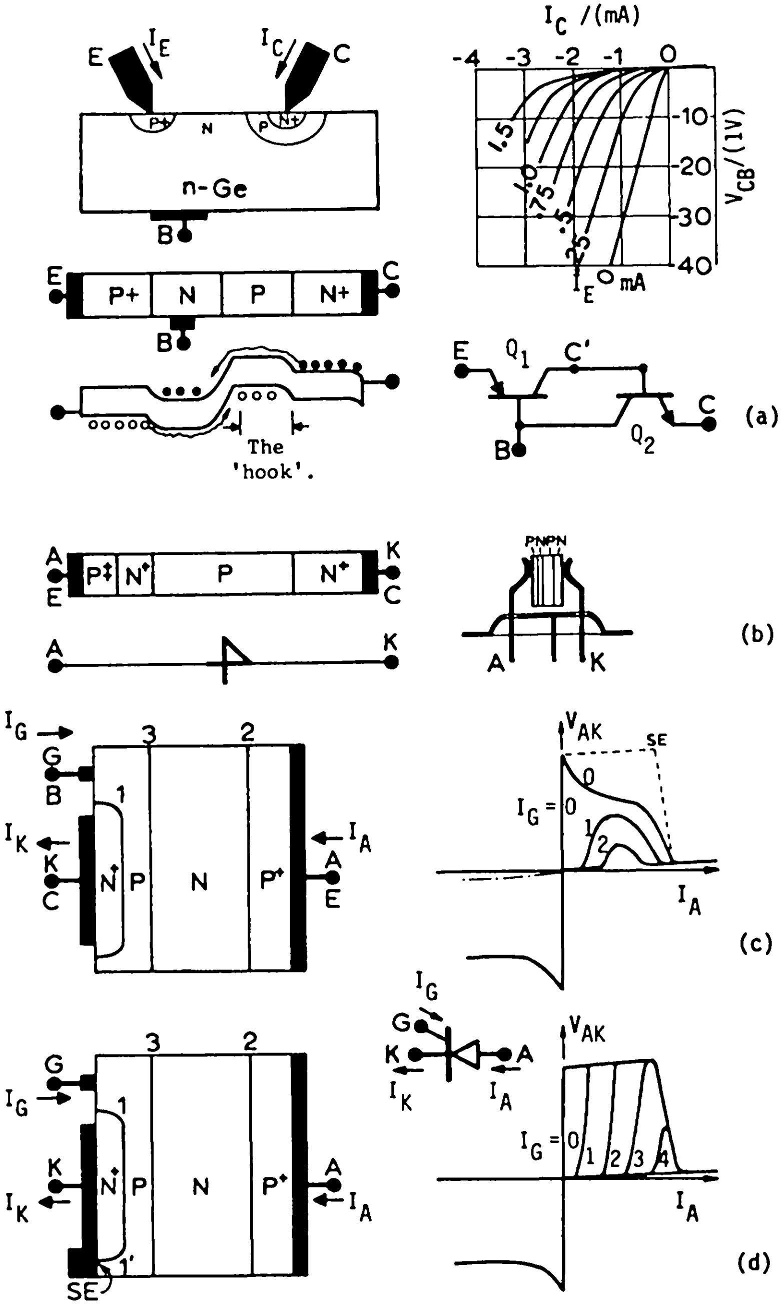


Fig.780.1 The voltage-current (V-I) characteristic of a typical four layer silicon diode manufactured by the Shockley Transistor Corporation in 1958. (a) Oscilloscope trace at 60Hz. (b) Expanded oscilloscope trace of the hysteresis loop. (c) Labeled V-I showing the notation convention. (d) Circuit symbol for 4-layer diode and triodes (SCR).

The lack of a control electrode and terminal has prevented the use of the bistable solid-state diode and the gas-filled diode tubes or cells in switching and control applications. A third electrode or terminal was added which has resulted in highly successful and widely used power control devices, such as the thyratron in gas-filled vacuum tubes and the 4-layer 3-terminal p/n/p/n solid-state triode known as the **thyristor** or the **silicon controlled rectifier** (SCR) which has replaced the thyratron tube. In this section, 780, and sections 781 and 782, the history and the electrical characteristics of the 4-layer p/n/p/n diode and triodes will be described. The MOS-SCR is described in section 783. The deleterious parasitic p/n/p/n that causes latch-up in CMOS and its avoidance are discussed in section 784.

We shall first define the notations and symbols using the d.c. current-voltage characteristics shown in Figs.780.1(a)-(d) from the historically first mass-produced silicon double-diffused silicon p++/n+/p/n+ 4-layer diode manufactured by the Shockley Transistor Corporation in 1958. Part (a) is the V versus I (V-I) characteristic from a transistor curve tracer. Part (b) is an enlarged view showing the hysteresis loop near knee or valley. Part (c) is the enlarged V-I trace with the labels used in today's thyristors or silicon controlled rectifier (SCR) (3-terminal 4-layer p/n/p/n devices). The same symbols can also be used for the 4-layer diode by letting the gate (base) current be zero. The forward peak voltage, $V_{F\text{-}PEAK}$, is called the forward breakover voltage in commercial data sheets and denoted by V_{BO}. The pioneer 4-layer diode and SCR researchers have also called this the forward breakdown voltage, V_B (which confuses with the junction barrier voltage); forward switching voltage or just the switching voltage, V_S (which confuses with the source and sustaining or holding/latching voltages; forward turn-on voltage, V_{TON}; and the forward firing voltage, $V_{F\text{-}fire}$ (from gas-discharge tube). The current at the peak voltage $I_{F\text{-}PEAK}$ is called the forward breakover, switching, turn-on or firing current ($I_{F\text{-}BO}$, I_{BO}, I_S, I_{TON} or $I_{F\text{-}fire}$). I_L and V_L are known as the **latching current and voltage**. The term 'latching current' has been used in power circuits and devices such as power switchgears, power machineries, and power solid-state and tube devices. It is defined by IEEE as the **making current** during the closing operation in which the device latches. In 4-layer diode and thyristors or SCR, latching current is defined as the minimum principal current (current through the main terminals) required to maintain the SCR in the ON-state immediately after switching from the OFF-state to the ON-state has occurred and the triggering signal used to switch has been removed. I_H and V_H are known as the **holding current and voltage** whose symbol has been a universal consensus. The holding current is defined as the minimum principal current required to maintain the 4-layer diode (or thyristor) in the ON-state after the latching current has been reached and after the triggering signal has been removed. The reverse characteristics are similarly labeled as indicated in Fig.780.1(c). These 4-layer characteristics are very similar to the negative resistance characteristics of the 3-layer n/p/n bipolar junction transistor described in section 739. The addition of a p/n/p transistor (the 'hook' collector or second minority carrier emitter) increases the amplification or gain of the positive feedback loop, resulting in a much lower holding or valley voltage in the 4-layer device than in the 3-layer device.

The history of the 4-layer diode and triodes (SCR) is reviewed next. Companies were started in the mid-1950's and folded a decade later. They were started by the expectation of superior performance and huge application volume. They folded because of unforeseen application limitations. The singular undeniable success was the birth of the Silicon Valley from failed 4-layer diode production.


E
I_E
I_C
C
P+
N
P
N+
n-Ge
B
E
P+
N
P
N+
C
B
The 'hook'.
I_C /(mA)
-4 -3 -2 -1 0
-10
-20
-30
-40
V_CB/(1V)
1.5
1.0
.75
.5
.25
I_E 0 mA
E
Q1
C'
B
Q2
C
(a)
A
E
P‡
N+
P
N+
K
C
A
K
PNPN
A
K
(b)
I_G
G
B
3
2
1
I_K
K
C
N+
P
N
P+
I_A
A
E
V_AK
SE
I_G = 0
0
1
2
I_A
(c)
I_G
G
K
A
I_K
I_A
V_AK
I_G = 0
1
2
3
4
I_A
G
I_G
3
2
1
K
I_K
N+
P
N
P+
A
I_A
1'
SE
(d)

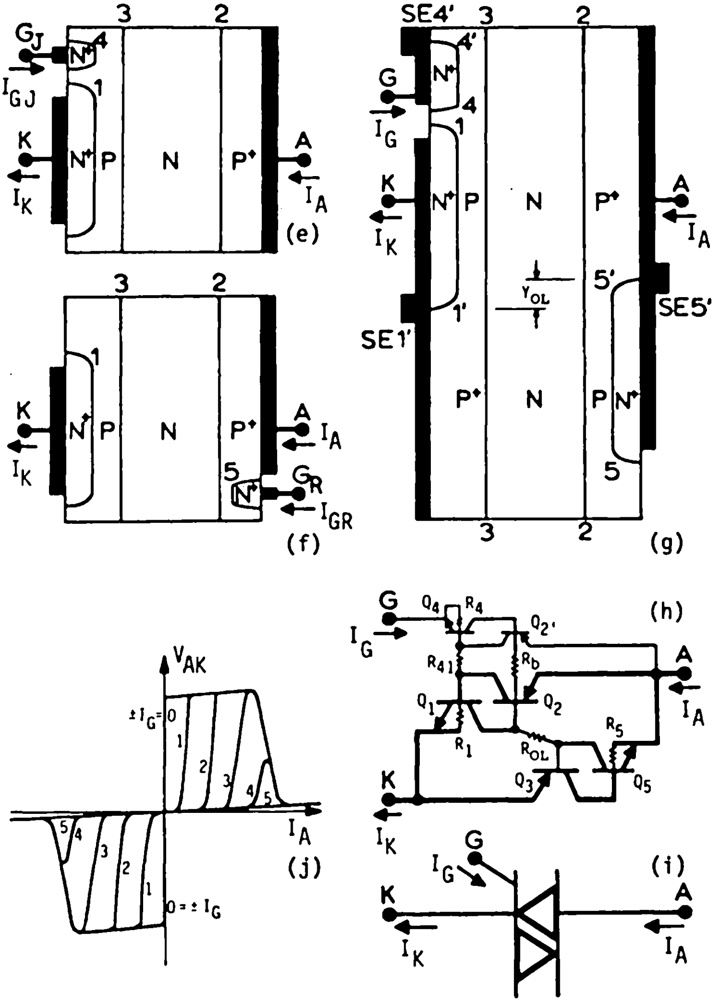

3
2
G_J
N^+
4
I_{GJ}
1
K
N^+
P
N
P^+
A
I_K
I_A
(e)
3
2
1
K
N^+
P
N
P^+
A
I_A
I_K
5
N^+
G_R
I_{GR}
(f)
SE4'
3
2
4'
G
N^+
4
I_G
1
K
N^+
P
N
P^+
A
I_K
I_A
Y_{OL}
5'
SE5'
1'
SE1'
P^+
N
P
N^+
5
3
2
(g)
G
Q_4
R_4
$Q_{2'}$
(h)
I_G
A
R_{41}
R_b
Q_1
Q_2
I_A
R_5
R_1
R_{OL}
K
Q_3
Q_5
I_K
V_{AK}
$\pm I_G = 0$
1
2
3
4
5
I_A
(j)
5
4
3
2
1
$0 = \pm I_G$
G
I_G
(i)
K
A
I_K
I_A

Fig.780.2 Evolution of Si p/n/p/n 4-layer diodes and triodes. (a) Shockley's 'hook' collector for greater than unity alpha in point-contact and p/n/p Ge BJT's and J.J. Ebers' complementary transistor simulation. (b) Shockley's 1958 4-layer p/n/p/n diode. (c) The 4-layer p/n/p/n triode or SCR with a ohmic gate contact to the the base layer next to the cathode. (d) Shorted emitter-base junction distributed-base-resistance SCR invented by Aldrich and Holonyak at the General Electric Co. (e) SCR with a junction-gate contact to the base layer next to the cathode invented by J.Moyson and F.E.Gentry of GE in 1961. (f) SCR with remote junction gate imbedded in the anode layer using the remote-base principle invented by R.N.Hall of GE in 1955. (g) The symmetrical bilateral 5-layer p/n/p/n/n SCR, the TRIAC, invented by F.E. Gentry of GE in 1965. (h) Equivalent-circuit of TRIAC. (i) Circuit symbol of TRIAC. (j) V-I characteristics of TRIAC.

Shockley's 'Hook' Collector Theory

Figure 780.2(a) illustrates the 'hook' collector model advanced by Shockley in 1950 to explain the large forward alpha ($\alpha_F > 2$) observed by Bardeen and Brattain in the point-contact or type-A transistor on n-Ge substrate or base in which they discovered the transistor effect in 1948. The type-A transistor and its I_C-V_{CB} curves are shown in the upper part of the figure. Shockley's p/n+ 'hook' collector in the p/n/p transistor is the origin of the p/n/p/n+ devices. In 1952, J. J. Ebers gave the first experimental circuit demonstration of greater-than-unity alpha by cascading a Ge p/n/p and a n/p/n transistor in a positive feedback loop. Ebers' circuit in Fig.780.2(a) consists of a CB input n/p/n transistor, Q_1, whose collector drives a CE p/n/p transistor, Q_2, in the emitter-follower configuration. In normal 2-T circuits, the base of Q_1 and collector of Q_2 would be grounded. In this case, they are not grounded but are tied together to simulate the internal feedback path of the p/n/p/n structure.

There are several mechanisms which can account for $\alpha_F > 1$ observed experimentally by Bardeen and Brattain in the type-A n-Ge phosphor-bronze point-contact as indicated by the I_C-V_{CB} curves in figure (a). The basic phenomenon of $\alpha_F > 1$ in a p/n/p BJT is as follow: the collector current consists of a hole and electron component, $I_C = I_{CP} + I_{CN}$, whose sum is greater than the forward emitter current which contains only a hole current, $I_E = I_{EP}$. The leakage currents are excluded, that is, the incremental emitter and collector currents are compared and calculated using

$$\alpha_f = (I_C - I_{CBO})/(I_E - I_{EBO}) \tag{780.1}$$

$$= \Delta I_C/\Delta I_E = (\Delta I_{CP} + \Delta I_{CN})/\Delta I_{EP} \tag{780.2}$$

$$\approx (M_p \alpha_b \gamma_e \Delta I_{EP} + \Delta I_{CN})/\Delta I_{EP} \tag{780.2A}$$

$$= M_p \alpha_b \gamma_e + (\Delta I_{CN}/\Delta I_{EP}). \tag{780.3}$$

where M_p is the multiplication factor of the emitter hole current that passes through the base/collector space-charge layer, for example, by interband impact generation discussed in section 384 and applied to the 3-layer p/n/p BJT in section 739. It is evident that this dissection gives two terms that can increase the alpha to more than unity: the multiplication term, M_p, due to interband impact generation of electron-hole pairs which is important at high reverse bias, and the collector electron current term, ΔI_{CN}, that is induced by hole space-charge from the injected hole current which can be important only at low (or nearly zero) collector/base voltages.

The original explanation given by Bardeen and Brattain was based on Schottky barrier lowering by the injected mobile space-charge (holes in n-Ge) from the term ($\Delta I_{CN}/\Delta I_{EP}$) in (780.3) because $\alpha_F > 1$ (as much as $\alpha_F > 3$) was observed at low V_{CB} where $M_p=1$. The device structure of the type-A n-Ge point-contact transistor is p+/n/???/m where ??? is Shockley's hook that was not included in the Bardeen-Brattain space-charge theory. The p+/n emitter/base point-contact junction was formed by a large positive current pulse. The positive forming current pushes the positive donor ions away from the emitter contact into the bulk of the n-Ge. Simultaneously, it attracts negative acceptor ions from the n-Ge bulk (presumably, the n-Ge has also some acceptor dopant impurity such as Al and Ga) towards the emitter contact to form the p+ region below the emitter contact surface as illustrated in Fig.780.2(a). The depletion of donor and accumulation of acceptor at the Ge surface under the emitter contact produces the p+/n junction which gave the observed good hole emitting property. In the space-charge theory, it was proposed that the Schottky barrier height of the n-Ge/metal collector (metal is a phosphor-bronze or tungsten wire of 0.5mil or 10μm point) is increased for holes but decreased for electrons (the total sum is fixed and equal to the Si energy gap) by the hole space charge from the emitter hole current that arrives at the collector space-charge layer. If this were true, then the lowered electron barrier height would increase the electron current of the collector junction and give a collector current greater than the emitter current or $\alpha_F>1$. This mechanism, if true, could be important only at very low collector reverse voltages when some electrons can still be injected into the collector metal from the n-Ge base. It cannot be important at large reverse collector bias, $V_{CB}<-4(kT/q)=-100mV$, because this electron injection current would drop to zero exponentially as $\exp(qV_{CB}/kT)$. Then, the electron current would be just the constant leakage component which does not depend on the barrier height or thickness of the semiconductor space-charge layer in the n-Ge. It only depends on the barrier height in the metal which is a fundamental property of the material that is independent of the space-charge density as indicated in section 562.

The fundamental error and failure of the barrier-lowering space-charge theory prompted Shockley in 1950 to propose the hook collector theory to account for $\alpha_F>1$ (as much as $\alpha_F>3$) at low voltages ($|V_{CB}|<0.9|BV_{CBO}|\sim100V$ for 10Ω-cm n-Ge) when $M_p\approx1$, and $\alpha_F>20$ in specially designed p/n/p/n in 1950. In the hook collector theory, the electron current is now injected by a second forward-biased electron emitting p/n+ junction which is in series with the collector/base n/p junction as illustrated by the cross-sectional views in Fig.780.2(a). This additional p/n+ junction was produced during forming using a current pulse applied to the collector/base point-contact which not only produces a good p/n collector junction but its heat also causes the phosphorus donor impurity in the phosphor bronze wire to diffuse into the p-collector layer to form a p/n+ junction which is the hook or the second emitter/base junction.

Shockley's Four-Layer Diode Venture

Figure 780.2(b) shows the four-layer diode manufactured in 1958 by the Shockley Transistor Corporation as a potential replacement for the mechanical switch (the Reed relay) used in the telephone switching offices. Its 1958 fabrication steps consisted of (1) growing a boron-doped p-Si crystal, (2) slicing the crystal into 0.012 inch (300μm) thick wafers, (3) mechanically lapping and chemically polishing to remove the saw damages and to thin down the wafer to 4-mil or 100μm, (4) predeposit a high-concentration phosphorus layer on both surfaces by exposing the wafer at 1000-1100C to phosphorus vapor from a heated P_2O_5 source kept at 210C to give 1mm P_2O_5 vapor in a two-zone (1100C and 210C) furnace, (5) diffusing the predeposited phosphorus into the p-Si in oxygen to give the two 25μm-thick n+layers, both covered also by an oxide film, (6) removing the oxide on the 'top' or better polished surface, (7) diffusing boron into the bared top surface also at 1100-1200C from a B_2O_3 powder source held at about the same temperature to give a p++ layer of about 15μm thick, (8) nickel-plate both surfaces, (9) alloy the nickel into the Si surface to give ohmic contact, (10) mask the wafer by 10-mil to 30-mil black-wax dots, (11) dice by chemical ($HF:HNO_3$) etch the masked wafer, and (12) mount each dice manually between two phosphor-bronze springs welded to the posts of a 3-lead TO-5 gold-plated transistor header as shown in Fig.780.2(b). Four-layer switching diodes, also trade marked as the **Shockley diode** with the symbol 4 were mass produced with breakover voltage from 10V to >300V and holding current from <1mA to >500mA. It is evident from the foregoing description that the production process was labor intensive and not well controlled. However, the main reason for its failure to achieve volume application as the telephone cross-point switch was the inability to reproduce the holding current accurately and to prevent transient turn-on by the capacitance current from a fast voltage pulse, CdV/dt. The theory in the following sections will analyze and describe the limitations. This failure was one of the main cause if not the true basic cause that created the dissension and difficult environment from undue profit-loss pressure which resulted in the departure of the key Shockley's staff who started the Fairchild Semiconductor Corporation to produce 3-layer 3-terminal n/p/n Si transistors rather than the hapless 4-layer 2-terminal n/p/n/p Si switching diodes. The spin-offs from Fairchild then started the growth of the Silicon Valley.

Silicon Controlled Rectifier (SCR)

The failure of the four-layer diode as a telephone cross-point switch did not deter the eventual success of the p/n/p/n device as a useful and widely used solid-state device. It was a simple extension of the fabrication steps using oxide-masking against diffusion and photolithography to provide a contact to one of the base layers. The incentive to develop a solid-state replacement of the gas-filled thyratron tube to control electrical power at 60Hz was recognized by the General Electric Company and Westinghouse. Many if not most of the 3-terminal SCR structures were invented and first manufactured by the General Electric Company during the first half of the 1960's. The evolution of the SCR is illustrated by the cross-sectional views of the SCR's in Figs.780.2(c) to (g).

Figure 780.2(c) shows the cross-sectional view that provides the contact for the base terminal as the input or control terminal. It also shows the d.c. voltage-current characteristic of three different devices: (solid curves) standard SCR with large reverse breakdown voltage, (dash-dot-dash) SCR with small or no reverse breakdown voltage using a parallel diode, and (--- labeled SE) SCR with shorted emitter to be described in the following subsection. Note that we have rotated the I-V (I vs V) curve by 90° to give the V-I (V vs I) curve in order to be able to talk about a negative resistance, two bistable voltage and current states, and the peak and valley voltages.

Note also that we have used a different labeling convention than those used in the pioneer research papers and early and recent power electronics books for a very simple reason: to conform with the IEEE convention of labeling two-port networks. Thus, the input lead and BJT are labeled by the numeral 1. The output lead and BJT are labeled by the numeral 2 which is in contrast to original literature which used numeral 3 that is inconsistent with the IEEE Standards on Symbols for the input/output terminals of a two-port network. The middle junction is labeled by the numeral 3 (for the 3rd junction) or by C (for collector junction when operated in the forward direction). Note also that the three terminals can be labeled by the ABC notation: A (for **Anode** or the Emitter), B (for the Base or Gate), and C (or K for **Cathode** or the collector of a p/n hook-collector or the electron emitter such as the hot-thermionic cathode in thyratron and vacuum tubes). But, to follow the convention used in the current power electronics textbooks and the IEEE Standards on Symbols for the thyristors, we shall use A for **Anode**, K for **Cathode**, and G for **Gate**. The standard SCR we shall describe and analyze will have a p-type Anode, a n-type cathode, and a p-type base next to the n-type cathode layer. Modification of the base structure and other parts of the diode can give a variety of desirable input or control characteristics. This input terminal will be labeled as Gate with a subscript, such as G_J for junction gate, G_S for shorted or shunted gate junction, G_R for remote gate, and G_O or G_I for oxide or insulated gate.

Return now to V-I characteristics of Fig.780.2(c). For the large reverse breakdown voltage SCR, the base lead is attached to the higher-alpha input transistor, 1 (in this case a n+/p/n), which has a lower emitter-base breakdown voltage and thinner base than the output p/n/p+ transistor, 2. Transistor 2 is essentially a symmetrical BJT whose emitter/base and collector/base junctions are nearly identical and have the same high reverse breakdown voltage. This base input location is particularly advantageous in most applications since the large breakdown voltage of the upper junction, 2, will isolate the input terminal, G, from any large reverse transient load voltage. If the base terminal is attached to the other base layer, then the full reverse load transient will appear in the input terminal. In addition, this high-alpha or high-beta base input terminal will provide a significant and useful turn-off gain, that is, the thyristor can be turned off by applying a negative gate current ($I_G<0$) whose magnitude is smaller than the anode or load current. A simple analysis will be made in section 782 to give the gate-turn-off gain formula and some curves.

Shorted-Emitter or Shunted-Emitter SCR

As just stated, the main reason for the unsuccessful application of the four-layer Shockley diode is the premature turn-on by a fast rising voltage which produces a capacitance current, Cdv/dt. The cause is the rise of alpha with current and the condition of switching or latching given by $\alpha_{F1}+\alpha_{F2}=1$ which is derived in the following sections, 781 and 782. Since the current dependence of the alpha comes from recombination in the emitter, emitter/base space-charge, and base layers, alpha is a very sensitive function of the recombination lifetimes in the three layers. Thus, the latch and holding currents are difficult to control in manufacturing because of the highly sensitive dependences of the condition $\alpha_{F1}+\alpha_{F2}=1$ on the density of the recombination centers. While this was the culprit that caused the failure of the four-layer diode as a commercial product, it also hinders the SCR because it uses the base current to turn-on the device and a base current transient is amplified by β_F to give β_F times larger transient in the anode or collector current which will falsely turn-on the SCR even more so. In addition, a fast anode current transient can also turn on the SCR.

Two remedies were invented which can be integrated monolithically. To alleviate the strong dependence on the recombination trap density in the base, the input emitter-base junction is shunted by its distributed base resistance via a short circuit metal on its far-side perimeter, 1', as indicated by SE in Fig.780.2(d). The resistive shunt causes α_1 to rise more sharply towards its maximum value at a higher and in a narrower range of emitter or cathode current. A fast base current will turn on only a small area of the nearby N+/P emitter/base junction, 1, and the rest of the SCR area is delayed by $r_{b'}\cdot C_{e'}$ from the lateral base resistance and the majority carrier or space-charge layer capacitance of the emitter/base (junction 1) and collector/base (junction 3) junctions. This base RC transmission line serves as a RC filter against the input transient. If the transient is a short and fast pulse, then only a small area of the p/n/p/n near the gate is turned on. This small area will be turned off quickly by the source resistance of the pulse generator to prevent the remaining and larger p/n/p/n from turning on. A similar short circuit can be used in the remote emitter/base junction, 2, to reduce a false turn-on by a fast load voltage transient on the output or anode terminal.

The shunt resistances lower α_{1F} and α_{2F} tremendously at low currents. When the anode current approaches I_H, they rise sharply to give $\alpha_{1F}+\alpha_{2F}=1$. Thus, the V-I curve is much squarish as indicated by the dashed curve in Fig.780.2(c) and the voltage spike near V_{BO} is eliminated. The spike is the cause of false turn-on by Cdv/dt at a voltage lower than the breakdown voltage of the collector junction, 3, because the small current width of the voltage spike is more easily overcome by a lower Cdv/dt.

Junction-Gate SCR

Figure 780.2(e) shows the n+/p-base junction gate geometry. It allows triggering or turning on the p/n/p/n SCR by a negative gate current in contrast to the positive triggering gate current applied to the ohmic contact to the p-base. The electrons injected by the n+ emitter gate diffuse through the p-base and reach the n-base to increase the base current of the p/n/p+ transistor, 2, and increases α_2 to cause $\alpha_1+\alpha_2=1$. It was first described by Moyson and Gentry of GE in 1961 and used as a part of the bilateral switch to be described.

Remote-Gate SCR

Figure 780.2(f) shows the remote n+/p junction gate geometry. It permits the SCR to be triggered from a terminal on the anode side without ohmic contact to the two base layers. It obviously works the same way as the junction gate just described, requiring a negative gate current which increases the base current of the p/n/p+ transistor and α_2 to turn on the SCR. It is based on the remote-base power transistor principle first proposed by R. N. Hall of GE in 1955.

Bilateral SCR

Figure 780.2(g) shows the combination of the three ideas just described, the shorted emitter, the junction gate, and the remote gate which is now as large as the cathode and shorted too. The cross section and the equivalent circuit show that it contains two reversely parallel large (main) SCR's and a small parallel trigger SCR. The remote gates are shorted to the respective anodes and act as the cathode of the other SCR. This 5-layer structure gives symmetrical SCR V-I characteristics shown in Fig.780.2(j). It can be triggered by either a positive or negative gate current in either the forward or the reverse mode, as indicated by the ± sign of $\pm I_G=0,1,2,\ldots$. Its generic acronym is TRIAC (TRIgger by AC).

781 Four-Layer PNPN Diode Characteristics

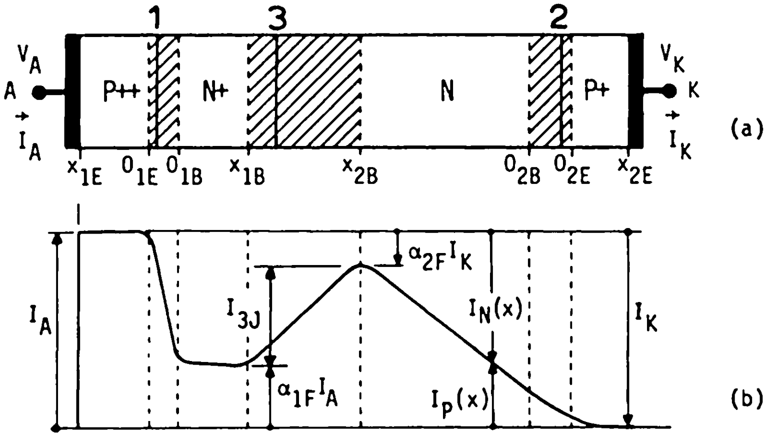
The negative resistance and bistable states in gas-filled tubes and the three-layer BJT and four-layer p/n/p/n solid-state diodes and triodes originate from a similar common principle: two mobile charge carrier species or current paths acting as the regenerative feedback path of each other to give loop gain greater than unity. The two carriers in the gas tube are the electrons and the neon ions (a small amount of argon, 0.1%, is added to lower the switching voltage). The two carriers in the solid-state diodes and triodes are the electrons and holes. The solid-state switch has one very significant feature: both species of charge carriers have high mobility and both can be readily injected in large quantity or high concentration at low voltages, i.e. by the two forward-biased p+/n and n+/p emitters. These are difficult to obtain in a gas tube. The operation consists of the following sequence. Electrons are generated by an external or internal means (a light or field ionization in the gas diode and the thermal generation of e-h pair by breaking the Si covalent bond in the SCR). The generated electrons are accelerated to high kinetic energies by a high voltage applied to the diode's two terminals. The energetic electrons collide with the neutral atoms in a gas and release their outer shell electrons or break the covalent bonds in a solid to generate more mobile electrons and holes. The positive ions or the holes are also accelerated to high kinetic energies by the applied voltage but in the opposite direction to that of the electrons. The energetic positive ions generate more electrons upon reaching the negatively biased electrode (the cathode) by impact release of metal electrons into the tube from the surface of the metal cathode. The energetic holes generate more electron-hole pairs in the semiconductor by breaking additional covalent bonds. There is a second regenerative path in the gas tube which is the light generated when some electrons are captured by ions. The photon from radiative capture can hit the cathode surface and release additional electrons via the photoelectric effect. These regenerative processes have a loop gain greater than unity because of the positive feedback from two energetic and charged particle species: the electrons and ions in the gas tube, and the electrons and holes in SCR. These regenerative processes would free all the bound or bond electrons if unrestrained to give tremendous (metallic) conductivity with electron density approaching that of a dense plasma. The regeneration will occur at localized high field regions to create a local avalanche and high current density even the total current is limited by external resistances. Thus, it is known as the avalanche multiplication process. In applications, a resistance in series with the diode or a limited load current will limit the device current to a value not much higher than I_H or I_L so it won't diverge. The terminal voltage and the internal electric field would drop to a low value just sufficient to give enough multiplication in order to sustain the current in the gas tube or enough injected currents from the three forward biased junctions to sustain the current.

The negative resistance and bistability phenomena in the 2-junction (p/n junction) 3-layer bipolar junction transistor was described and analyzed in section 739. The negative resistance and bistability were shown to be due to two regeneratively coupled positive feedback mechanisms: (i) increasing current due to interband impact generation of electron-hole pairs in the collector-base junction and (ii) increasing injected electron current (in n/p/n) or current gain, alpha, due to decreasing or lower rate of increase of recombination loss with increasing current density. The negative resistance (or decreasing voltage with increasing current) ceases and reverses back to positive resistance when the percentage of the injected electron current (relative to the total current) or the alpha decreases at higher total current densities due to decreasing emitter injection efficiency or increasing recombination rate in the base layer with increasing current density from the high injection level effect in the base layer while the heavily doped emitter layer is still at low injection level. In this and next section, 781 and 782, we shall extend the analysis of section 739 for the 2-junction 3-layer BJT to the 3-junction 4-layer bipolar junction diodes (2 terminals) and triodes (3 terminals). or SCR's (semiconductor controlled rectifiers).

D.C. Voltage-Current Characteristics of 4-Layer Diode

The d.c. characteristics of the 4-layer p++/n+/p/n+ diode can be analyzed using the electron and hole current diagram shown in Fig.781.1(b). By inspection, the current flowing in the collector junction (the middle junction labeled 3) is

$$I_3 = M_1\alpha_{1F}I_A + M_2\alpha_{2F}I_K + M_3I_{3J} \qquad (781.1)$$



Figs.781.1 A 4-layer diode. **(a)** Cross-sectional view. **(b)** D.C. hole and electron currents vs x.

The forward alpha's, α_{1F} and α_{2F}, were defined by (734.7B). An analytical approximation was derived in (739.6). Replacing I_C by I_E, this is

$$\alpha_F^{-1} = [1 + (j_E/j_B) + \sqrt{(I_{Eit}/I_E)} + (I_E/I_{EHL})]/\alpha_B \qquad (781.2)$$

$$= \alpha_{F\infty}^{-1}[1 + \sqrt{(I_{Eit'}/I_E)} + (I_E/I_{EHL'})] \qquad (781.3)$$

where

$$\alpha_{F\infty} = \alpha_B/[1 + (j_E/j_B)] \qquad (781.3A)$$

$$I_{Eit'} = I_{Eit}/[1 + (j_E/j_B)]^2 \qquad (781.3B)$$

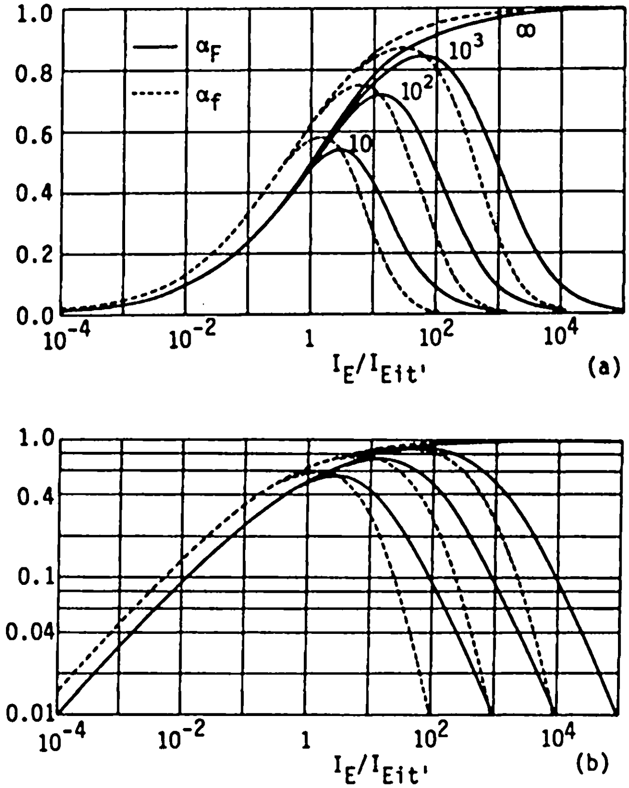
$$I_{EHL'} = [1 + (j_E/j_B)]I_{EHL} \qquad (781.3C)$$

Figures 781.2(a) and (b) on the next page illustrate the variation of the forward alpha with current in linear and semilogarithmic scales respectively.

I_{3J} is the current flowing in junction 3 due to recombination-generation in its space-charge layer and the two adjacent quasi-neutral base layers. The explicit expressions were given by (733.20) and (733.21A) to (733.21F) which can be grouped into two terms:

$$M_3I_{3J} = +M_{10}I_{10}[1 - \exp(qV_3/kT)] + M_{20}I_{20}[1 - \exp(qV_3/2kT)]. \qquad (781.4)$$

V_3 is the voltage measured from the p- to n-side of the collector junction.



Figs.781.2 Variation of the normalized d.c. and low-frequency a.c. or differential current gains, $\alpha_F/\alpha_{F\infty}$ (solid) and $\alpha_f/\alpha_{F\infty}$ (dash), with normalized emitter current. The parameter is $I_{EHL'}/I_{Eit'}$. (a) Semilog plot. (b) Log-log plot.

For the two-terminal diode, there is no base lead to give additional current flowing into the device, thus, $I_3 = I_A = I_K$ and (781.1) reduces to

$$I_A = M_1 \alpha_{P1} I_A + M_2 \alpha_{P2} I_A + M_3 I_{3J} \tag{781.5A}$$

or

$$I_A = M_3 I_{3J}/(1 - M_1 \alpha_{1P} - M_2 \alpha_{2P}). \tag{781.5B}$$

M_1 and M_2 are the multiplication factors of the currents injected into the space-charge layer of the central junction due to interband impact generation of electron-hole pairs. M_3 is the current multiplication factor of electrons or holes generated in the two quasi-neutral base layers that are swept into the space-charge layer, and electrons and holes generated inside the space-charge layer. These multiplication factors were described in detail in section 536. A simple formula of M was derived when the generation rate of electron-hole pairs by electron impact is equal to that of hole impact. This was given by (536.16) and (536.17) when $\alpha_n=\alpha_p=\alpha(E)=\alpha_0\exp[-b/|E(x)|]$ where α_0 and b are given by the fundamental parameters that describe the energy-momentum conservation of the interband electron-generation mechanism in the materials. The formula was given by (536.17) and is

$$M^{-1} = 1 - \int_{x_{1B}}^{x_{2B}} \alpha(x)dx = 1 - \int_{x_{1B}}^{x_{2B}} \alpha_0\exp[-b/|E(x)|]dx \qquad (781.6)$$

$$\approx 1 - (V_3/V_{3B})^n. \qquad (781.6A)$$

Figure 781.3 illustrates $\int\alpha(x)dx=1-M^{-1}$ versus V_3/V_{3B} in linearly graded Si p/n junctions with V_{3B}=20V, 60V and 100V. It shows that n varies from 1 to 3 at V_{3B}=100V and 1 to 4 at V_{3B}=20V. Similar variations of n occur also in abrupt p+/n and n+/p junctions.

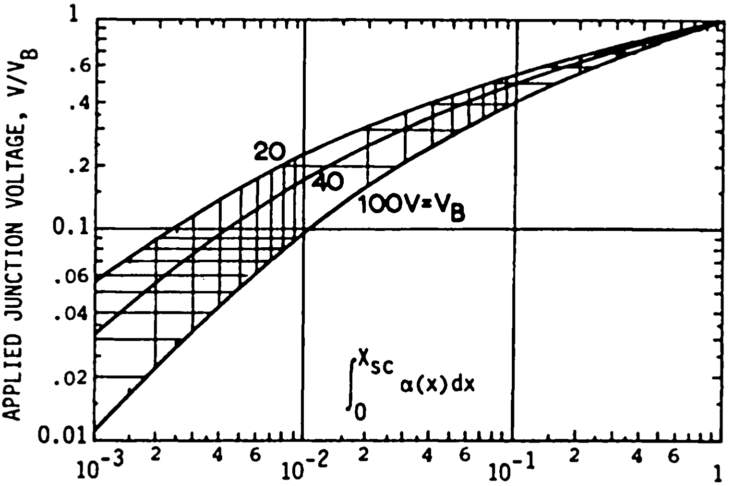


Fig.781.3 Illustration of the integrated impact generation coefficient, $\int\alpha(x)dx=1-M^{-1}$, vs the normalized applied reverse junction voltage, V_3/V_{3B}, in linear graded Si p/n junction assuming equal ionization rate by electron and hole impact, $\alpha_n(x)=\alpha_p(x)=\alpha(x)$. Subscript 3 is dropped.

Under the simple assumption of $\alpha_n(x)=\alpha_p(x)=\alpha(x)$, the current-voltage equation of (781.5B) becomes

$$I_A = MI_{3J}/[1 - M(\alpha_{1F}+\alpha_{2F})] = I_{3J}/[M^{-1} - (\alpha_{1F}+\alpha_{2F})] \qquad (781.7)$$

or

$$1 - M^{-1} = \int\alpha(x)dx$$

$$= (V_3/V_{3B})^n = 1 - \alpha_{1F} - \alpha_{2F} - (I_{3J}/I_A). \qquad (781.8)$$

For a given I_A, V_3 can be computed from (781.8) using $\alpha_F(I_A)$ from (781.3). Figure 781.4 shows that the computed V-I curve (dashed) compares well with experiment (solid) with $V_{BO}=15V$ and $I_H=2.5mA$.

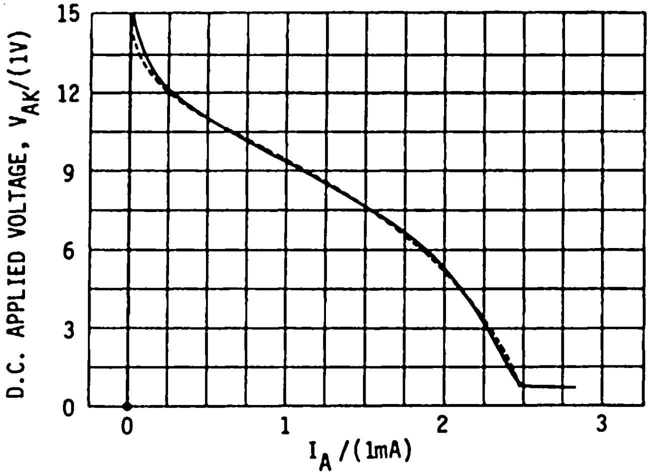


Fig.781.4 Theoretical-experimental comparison of a Si p++/n+/p/n+ 4-layer diode.

Physics of the 4-Layer Diode V-I Characteristics

The V-I characteristic can be divided into three regions in order to gain a detailed insight of device physics. In the low current positive resistance range below V_{BO} and I_{BO} [not visible in Fig.781.4 but clear in Fig.782.1(a)], the alpha's are very low because of the very low current which is essentially given by the generation current of the middle or collector junction, 3, under reverse bias. As the voltage increases towards breakdown, the current increases due to multiplication in the high-field space-charge layer of the middle junction, MI_{3J} in the numerator of (781.7). When the current multiplication becomes very large, the alpha's start to increase toward their maximum value and the current increases further due to the denominator $1-M(\alpha_{1F}+\alpha_{2F})$ of (781.7). This positive feedback causes the voltage to decrease in order to decrease M so that the current reaches a deterministic value. The decreasing voltage with increasing current gives the negative resistance region which is the second region. Since the current in this region is much larger than the leakage current, I_{3J}/I_A can be dropped in (781.8) and the negative resistance characteristic is accurately described by

$$M(\alpha_{1F} + \alpha_{2F}) = 1 \tag{781.9A}$$

or

$$1 - M^{-1} = (V_3/V_{3B})^n = 1 - (\alpha_{1F} + \alpha_{2F}). \tag{781.9B}$$

Thus, the negative resistance characteristic is given by the condition of **unity total alpha** where the total alpha is defined as $\alpha \equiv M(\alpha_{1F}+\alpha_{2F})$. The third region is the high-current/low-voltage positive resistance region where $M \approx 1$ because of the low voltage across the middle junction, 3. From (781.7), the V-I characteristic is

$$I_A = I_{3J}/(1 - \alpha_{1F} - \alpha_{2F}). \qquad (781.10)$$

Since $\alpha_{1F}+\alpha_{2F}$ will exceed unity as the current increases, I_{3J} must become negative in order to give a positive I_A. From (781.4) with $M=1$, I_{3J} becomes

$$I_{3J} = + I_{10}[1- \exp(qV_3/kT)] + I_{20}[1 - \exp(qV_3/2kT)]. \qquad (781.11)$$

which shows that if $I_{3J} < 1$, V_3 must become positive or the middle junction is forward biased and injecting minority carriers into the two adjacent base layers. This is similar to the BJT driven into the saturation range by a large base current, causing its collector/base junction to be forward biased. Thus, to turn-off the diode, the additional minority carrier charges must be extracted by base recombination which is slow with time constant τ_B, shorted emitter/base junction, or a reverse gate current, I_R, which speeds up the turn-on by the current ratio of I_R/I_F as in the BJT or the p/n junction diode. The voltage-current characteristic in the high-current low-voltage 'on' state can again be calculated for a given diode current, I_A, since the three junction voltages are explicitly related to the diode or junction current. If one of the two emitter junctions and the middle junction are identical, such as the Shockley four-layer diode produced by the double diffusion process described early, then the on-state V-I curve is just the V-I curve of the dissimilar junction or the emitter/base junction of the high-alpha transistor, p++/n+ in Fig.780.2(b), because the other two identical junctions cancel exactly. The holding current can be defined as the current at which the voltage across the central junction is zero or

$$\alpha_{1F}(I_H) + \alpha_{2F}(I_H) = 1. \qquad \textbf{Holding Condition} \quad (781.12)$$

The holding voltage is then approximately given by

$$I_H = I_{10}[\exp(qV_H/kT) - 1] + I_{20}[\exp(qV_H/2kT) - 1] \qquad (781.12A)$$

$$\approx I_{10}\exp(qV_H/kT) + I_{20}\exp(qV_H/2kT). \qquad (781.12B)$$

A general analytical result can also be obtained for the switching (breakover) and holding conditions because the voltage reaches a maximum at switching and a minimum at holding, thus, $dV_A/dI=0$ at these two conditions. This can be applied to the general V-I equation, (781.7). Taking the derivative with respect to I_A and using $dM/dI_A=(dM/dV)(dV/dI)=0$, then

$$\alpha_{1f} + \alpha_{2f} = M^{-1} \qquad (781.13)$$

where α_f is the a.c. low-frequency or differential alpha defined by

$$\alpha_f = dI_C/d_B = \alpha_F + \partial\alpha_F/\partial\log_e I_B$$

$$= \alpha_F(1 + \partial\log_e\alpha_F/\alpha\log_e I_B). \qquad (781.14)$$

The general result, (781.13), has a very significant difference from that given in the literature. In the literature, the factor M is inadvertently omitted because multiplication of the leakage current in the middle collector junction, 3, was overlooked. It is evident that unless α_f or α_F is very high at very low currents, M must be quite large in order to satisfy (781.13). In Si transistors, alpha is very low at low currents, thus, M must be quite large or the breakover voltage must be rather close to the breakdown voltage of the middle junction, 3. At low currents, α_F is proportional to a power of the current, I^m, ($m=0.5$ in the simple SNS recombination model used in section 738), thus, $\alpha_f=(1+m)\alpha_F=(3/2)\alpha_F$ and the breakover condition, (781.13), becomes

$$M^{-1} = \alpha_{1f} + \alpha_{2f} \qquad \textbf{Breakover Condition} \quad (781.15)$$

$$\propto (1+m)(\alpha_{1F} + \alpha_{2F}) \qquad (\text{if } \alpha_F \propto I_A{}^m) \qquad (781.15A)$$

or

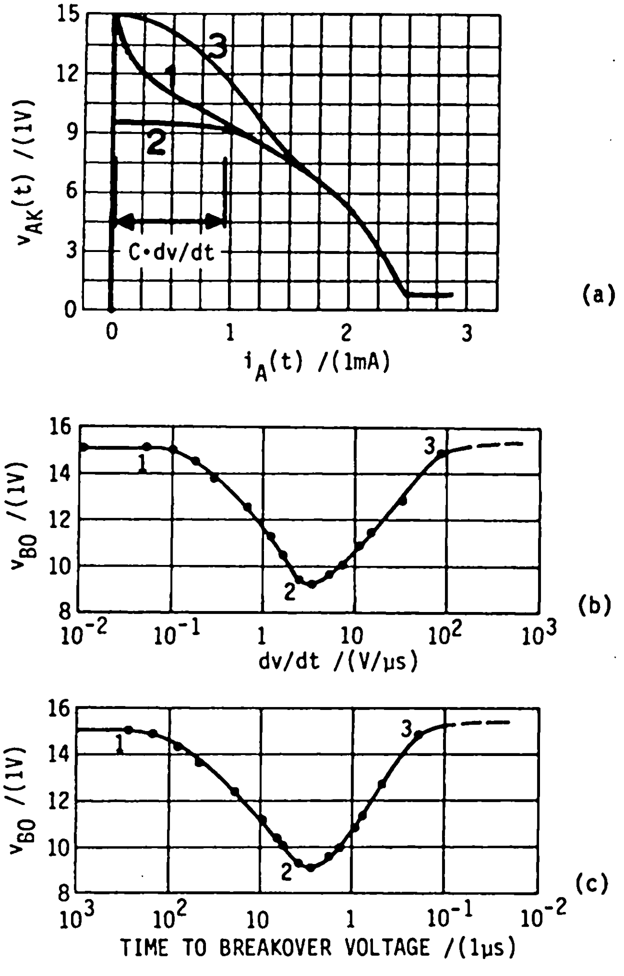
$$\alpha_{1F} + \alpha_{2F} = [(1+m)M]^{-1} \propto 1/(1+m) \quad (\text{if } M \propto 1) \qquad (781.15B)$$

$$\propto 2/3 \qquad (\text{if } m=1/2). \qquad (781.15C)$$

Combining the holding condition given by (781.12) and the breakover condition given above, we see that the negative resistance range starts at $\alpha_{1F} + \alpha_{2F} = 2/3 = 0.667$ and ends at $\alpha_{1F} + \alpha_{2F} = 1$.

Switching Transient in 4-Layer Diode

The 4-layer diode can be prematurely turned on at a voltage lower than V_{BO} by a fast voltage transient due to the displacement current passing through the capacitance of the emitter and collector junction space-charge layers, $C_j dv/dt$. The symptom is that the **dynamic breakover voltage**, $v_{BO}(dv/dt>0)$, is lower than the static value, $v_{BO}(dv/dt\to0)=V_{BO}$. This capacitance or displacement current is the majority carrier base current which has a very small time constant, about 10ps estimated in section 752 for the phase I switch-on transient in the CB BJT configuration. In contrast, the capacitance currents from charge storage in the two base and two emitter quasi-neutral layers are minority carrier diffusion currents which experience diffusion delays through the quasi-neutral layers of $t_{B1}=X_{B1}{}^2/2D_{B1}$, t_{B2}, t_{E1}, and t_{E2}. Their (majority and minority carriers) effects on the dynamic breakover voltage and premature turn-on are opposite. This difference is illustrated in Figs.781.5(a)-(c). Figure (a) shows the v(t)-i(t) loci for the three dv/dt values labeled 1, 2, and 3 in figure (b) which gives v_{BO} vs dv/dt measured on the Shockley 4-layer diode shown in Fig.781.4. Figure (c) gives v_{BO} as a function of the duration necessary to reach the v_{BO} at various ramp rates dv/dt of a linear ramp trigger. These figures show that if dv/dt is very small, v(t) will rise to the static breakover value, $V_{BO}=15V$.



Figs.781.5 The transient response of a 4-layer p/n/p/n diode. **(a)** The v(t)-i(t) loci at three increasingly higher dv/dt. **(b)** The dynamic breakover voltage as a function of dv/dt. **(c)** The dynamic breakover voltage as a function of the duration to reach the dynamic breakover voltage using a linear voltage ramp as the trigger.

When dv/dt is increased, the capacitance current Cdv/dt will increase the alpha and lower the breakover voltage to $v_{BO}(dv/dt>>0) < V_{BO}$. For example, curve 2 corresponds to a Cdv/dt=1mA or about $I_H/2$ and the breakover voltage drops to 9V from d.c. value of 15V. When dv/dt increases further to about dv/dt=100V/μs, v_{BO} rises back to 15V because during the 150ns duration to reach 15V [computed from $v_{BO}/(dv/dt)$ = (15V)/(100V/μs) = 150ns] most of the minority carriers injected by the emitters have not arrived at the central junction, 3, to start multiplication owing to the diffusion delay in the two base layers. At still faster voltage rise or larger dv/dt, v_{BO} will reach the maximum given by $V_{EB1}+V_{3B}+V_{EB2}$ where V_{EB1} and V_{EB2} are a fraction of the injection threshold value [$V_{it}\simeq 0.26V$ from (535.7)] because none of the minority carriers has arrived at the center collector junction owing to diffusion delay in the base layers. The current is limited by the source resistance of the dv/dt ramp generator.

The reduction of the breakover voltage due to the majority carrier capacitance current can be easily derived by extending the static V-I equation. Denote the displacement or capacitance current by $i_{DIS}(t)$, (781.1) is modified to give

$$I_3 = M_1\alpha_{1P}I_A + M_2\alpha_{2P}I_K + M_3I_{3J} + i_{DIS}(t). \qquad (781.16)$$

Assuming $\alpha_n=\alpha_p$ so $M_1=M_2=M_3=M$ and using $I_3=I_A=I_K$, (781.16) becomes

$$I_A = (I_{3J} + M^{-1}i_{DIS})/[M^{-1} - (\alpha_{1P}+\alpha_{2P})]. \qquad (781.17)$$

Using the breakover condition, $dV/dI_A=0$, then $d(781.17)/dI_A$ gives

$$\alpha_{1f} + \alpha_{2f} = 1/M. \qquad (781.18)$$

This is identical to (781.13) which was derived without the displacement or capacitance current. The effect of displacement current on v_{BO} comes from the current variation of the a.c. α_{1f} and α_{2f} via the d.c. α_{1F} and α_{2F} or (781.4) which is given by (781.2) and (781.3) because I_E is now given by $I_A+i_{DIS}(t)$. The breakover voltage is lowered because $\alpha_{f1}+\alpha_{f2}$ is increased to unity by the presence of the capacitance/displacement current so that it is not necessary to use $M>1$ by having a high voltage at the middle junction, V_3, to reach the $M(\alpha_{f1}+\alpha_{f2})=1$ breakover condition. The validity of the algebraic addition of the displacement current $i_{DIS}(t)$ to the static anode current, I_A, is based on the quasi-static approximation in which the diffusion delay in the two base layers are neglected. Thus, the theory will not predict the rise of V_{BO} at higher dv/dt. It should also be noted that narrower current spike at low currents in the V-I curve is more easily wiped out by a slower voltage ramp, resulting in a higher sensitivity to voltage transients. Thus, the ideal V-I would be a rectangle as suggested by the dash curve in Fig.780.2(d).

782 PNPN Triode (SCR) Characteristics

As indicated in Figs.780.2(c)-(g), the p/n/p/n can be turned on by a signal applied to the base layer. It can also be turned off by drawing a current out of the base layer. Thus, the attachment of a terminal to the base layer next to the cathode will facilitate both the turn-on and turn-off of the p/n/p/n switch. This 3-terminal 4-layer transistor is known as the semiconductor (or silicon) controlled rectifier: its rectification properties are controlled by the base terminal. There are several structural additions discussed in Figs.780.2(c)-(g) which give unique and new properties. These will now be analyzed.

Basic SCR

From Fig.780.2(c), it is evident that $I_K=I_A+I_G$ where $I_3=I_A$ and I_G is the current flowing into the p-base of the input n/p/n transistor. Substituting these into the 4-layer diode equation, (781.1), which is repeated below with I_K and I_A interchanged so that 1 refers to input and 2 refers to output,

$$I_3 = M_1\alpha_{1F}I_K + M_2\alpha_{2F}I_A + M_3I_{3J}, \qquad (782.1)$$

then the triode or SCR d.c. equation is

$$I_A = (M_3I_{3J} + M_1\alpha_{1F}I_G)/(1 - M_1\alpha_{1F} - M_2\alpha_{2F}). \qquad (782.2)$$

Making the simplifying assumption $\alpha_n=\alpha_p$ so that $M_1=M_2=M_3=M$, then

$$I_A = (I_{3J} + \alpha_{1F}I_G)/(M^{-1} - \alpha_{1F} - \alpha_{2F}) \qquad (782.3)$$

$$\simeq \quad \alpha_{1F}I_G/(1 - \alpha_{1F} - \alpha_{2F}) \quad \text{(Near Breakover)} \qquad (782.3A)$$

where (782.3A) is valid when $\alpha_{1F}I_G >> I_{3J}$ which is generally the case during gate triggering to turn on the SCR.

The turn-on condition is again obtained from $dV/dI_A=0$ and given by (781.13), just like the case of having a fast dv/dt capacitance current in the anode or load line derived in (781.18),

$$\alpha_{1f} + \alpha_{2f} = 1/M \simeq 1. \qquad \text{(When } \alpha_{1F}I_G>>I_{3J} \text{ so } M\simeq 1.) \qquad (782.4)$$

The breakover voltage is lowered by the gate current I_G through its influence on increasing the I_A tremendously via (782.3A). The larger I_A in turn increases $\alpha_{1f}(I_E=I_K=I_A+I_G)$ and $\alpha_{2f}(I_E=I_A)$ via (781.2).

A delay will be encountered when turning on the SCR from the base terminal to reach a forward current I_F. This delay is just like the delay encountered in turning on a BJT in the CE configuration discussed in section 753. It consists of two parts, an initial short delay to reach the breakover condition which is much shorter because the base charge required is much smaller, and a longer delay to build up the large excess minority carrier charge stored in the base layers from injection by all three forward biased junctions in order to give the large forward current I_F. In SCR, the base of the output transistor is usually much thicker in order to give a high breakdown voltage so the stored charge is the larger of the two. Thus, the turn-on delay to reach $0.9I_F$ is roughly given by (753.6C) or $2.3\tau_B$. In practice, there is also a lateral delay due to $r_{b'}\cdot(C_c+C_e)$ because practical SCR's are two-dimensional devices to make use of the lateral delay to counter the turn-on by fast rising and transient load voltages and turn-off by fast transient load currents that may appear at the anode terminal.

The SCR can also be turned off by drawing a portion of the anode current out of the gate or base terminal to reduce the middle junction voltage to zero. Using the transistor gain mechanism, the SCR can be designed to make this base turn-off current only a fraction of the anode current, thus, giving a turn-off gain. To obtain an analytical formula, we use the general V-I equation (782.3). In the on region, all three junctions are forward biased as explained in (781.10) and (781.11), thus $M \approx 1$ and $I_{3J} < 0$, which are substituted into (782.3) to give

$$I_A = (|I_{3J}| - \alpha_{1F} I_G)/(\alpha_{1F} + \alpha_{2F} - 1). \tag{782.5}$$

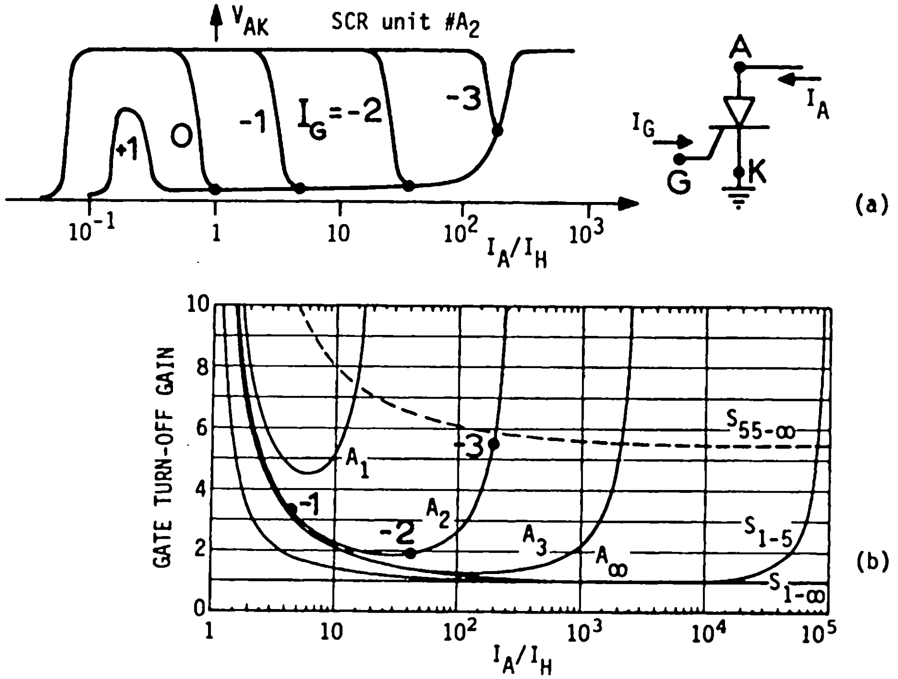


Fig.782.1 The gate turn-off characteristics of a SCR. (a) D.C. V_{AK} vs I_A curves. (b) The gate turn-off gain, $I_A/(-I_G)$. Curves A_n are for $\alpha_1 = 0.6, \alpha_{2\infty} = 1.0$, $I_{BCHL'}/I_{BCH'} = 10^n$. Curves $S_{1\text{-}n}$ are for $\alpha_{1F} = \alpha_{2F}$, $\alpha_{1\infty} = \alpha_{2\alpha} = 1.0$. Curve $S_{55\text{-}\infty}$ is for $\alpha_{1F} = \alpha_{2F}$, $\alpha_{1\infty} = \alpha_{2\alpha} = 0.55$.

Thus, drawing a current out of the base gate terminal ($I_G < 0$) would reduce the anode current to less than I_H or the forward bias on the middle junction to zero to give $I_{3J}=0$. This would turn off the SCR. Supposing that this gate current is $I_G=-I_{G\text{-}OFF}$ and let $V_3=0$ or $I_{3J}=0$, then the gate turn-off gain from (782.3) is

$$G_{GTO} = I_A/I_{G\text{-}OFF} = \alpha_{1F}/(\alpha_{1F} + \alpha_{2F} - 1). \tag{782.6}$$

It is evident that to get large gate turn-off gain over a range of high anode current, the sum of the maximum alpha should exceed unity only slightly. For example, a gain of 10 is obtained if $\alpha_{1F}=1.0$ and $\alpha_{2F}=0.1$. If both BJTs have nearly unity maximum alpha, then the gate turn-off gain is 1 or there is no gain. Figure 782.1(b) shows the calculated gate turn-off gain versus normalized anode current using the alpha variation with current given by (781.2) and shown in Figs.781.2(a) and (b). These curves show substantial gate turn-off gain if the SCR is operated near the holding current. In practice, the gain curve is more rectangular due to the sharper rise and fall of α with current due to the emitter shunt resistances. The practical gain is not high because the control of α_{1F} and α_{2F} to make $\alpha_{1F}+\alpha_{2F}$ nearly 1 is difficult.

The turn-off delay can also be estimated using the turn-off delay of a p/n junction analyzed in section 553. Since all three junctions are forward biased and usually one of the base is thick compared with the diffusion length, the turn-off delay is essentially all due to the recombination delay of the excess minority carriers in the thick base layer. Thus, using the analogy to (553.3), the time to reduce the stored charge from a forward current value I_F to the holding current I_H is

$$t_{OFF} = \tau_B \log_e(I_F/I_H) \tag{782.7}$$

where τ_B is the recombination lifetime in the low-alpha thick-base layer which is usually the high resistivity bulk silicon. Thus, in SCR, doping with a recombination impurity such as gold is a necessity to speed up turn-off of the rectifier. For power control applications, the shorter turn-off time would mean lower energy loss and higher efficiency. However, higher recombination centers and lower recombination lifetime, τ_B, would also mean higher leakage current whose power dissipation must be taken into account.

Shorted-Emitter or Shunted-Emitter SCR

The breakover voltage and hold current are difficult to control because they are very sensitively dependent on the variation of the alpha with current. To reduce this sensitivity, a metallic short circuit at the remote surface of the emitter-base junction was invented. In addition, the remote short circuit will also reduce premature turn-on by the capacitance current due to a fast voltage ramp because of the lateral $r_{b'}\cdot C_j$ delay of the trigger voltage. Figure 780.2(d) shows the shorted emitter SCR. The V-I characteristics is given by a modified (782.3)

$$I_A = (I_{3J} + \alpha_{1F'}I_G)/(M^{-1} - \alpha_{1F'} - \alpha_{2F}) \tag{782.8}$$

where

$$\alpha_{1P'} = \alpha_{1P}/(1 + I_{SHUNT}/I_{1J}) = \alpha_{1P}/[1 + (V_1/r_s I_{1J})] \qquad (782.8A)$$

and

$$I_{1J} = I_{10}[\exp(qV_1/kT)-1] + I_{20}[\exp(qV_1/2kT)-1]. \qquad (782.8B)$$

The distributed base resistance is approximated by a lumped shunt resistance, r_s, which is approximately $\frac{1}{2} r_{b'}$. The shunt resistance reduces the low-current alpha of the high-alpha input transistor sharply at a designed current, V_{EB1}/r_s, where $V_{EB1} \approx 0.5$V. Thus, the rise of alpha with current can be better controlled and the V-I curves are closer to a rectangle sketched in Fig.780.2(d) and do not have the voltage spike at low-current. This improves its resistance to premature turn-on by a fast voltage transient.

Junction-Gate SCR

The n/p junction-gate SCR is shown in Fig.780.2(e). It is invented to give a negative gate triggering current in order to provide bipolar triggering capability when used with the ohmic contact gate of Fig.780.2(c) which employs a positive gate triggering current. A negative current through the n+/p junction, 4, will inject electrons from the n+ emitter through the p-base to the n-base of the main n+/p/n/p+ area. This base current of transistor 2 will increase α_{2F} and causes $\alpha_{1F}+\alpha_{2F}$ to approach 1 and the SCR to switch. The V-I equation is given by

$$I_A = (I_{3J} + \alpha_{2P}\alpha_5 I_G)/(M^{-1} - \alpha_{1P} - \alpha_{2P}). \qquad (782.9)$$

Remote-Gate SCR

This was shown in Fig.780.2(f). It is a five layer device with the addition of the remote n+/p+ gate on the anode p+layer. Again a negative current is applied to the n+ gate to inject electrons into the p+ emitter which then diffuse to the n-base of the p+/n. In the n-base, these electrons increase the base current of the p+/n/p+ transistor, causing the SCR to turn on. The V-I equation can be immediately written as:

$$I_A = (I_{3J} + \alpha_{2P}\alpha_{4P} I_G)/(M^{-1} - \alpha_{1P} - \alpha_{2P}). \qquad (782.10)$$

Bilateral SCR

This was shown in Fig.780.2(g). It combines the three variations just described, the shorted emitter, the junction gate, and the remote gate that is shorted when serving as the emitter for the reverse SCR. These combinations give a symmetrical SCR that can be triggered by either a positive or a negative current and operate in either current directions. The V-I equation is more complex but can be readily written down using (782.8)-(782.10).

783 MOS-SCR

Figure 783.1(a) shows a monolithic MOS-SCR using an insulated or oxide gate to control the base current of the input n/p/n BJT. This structure was first employed by Sah to study the effect of surface recombination and channel on the BJT alpha and beta in 1961 which was described in section 738 and references [733.3] and [733.4] or [737.3] as well as [600.1]. The original structure used in the surface recombination and surface channel experiments contained only a shunt pMOST across the emitter/base junction to control the alpha or beta of the BJT. In Fig.783.1(a) a series nMOST is added to turn on the SRC. This also makes the input a CMOS with common emitter output as shown in circuit diagram given in Fig.783.1(b). The advantage of the CMOS gate is obviously the high input impedance and extremely large power gain. However, for controlling large load current, the pMOST and V_{SS} must have a substantial current carrying capability to turn off the SCR unless the gate turn-off gain is very high which is hard to reproduce as discussed previously. Thus, the CMOS gate is useful only when the load current is not extremely high, such as 1kA.

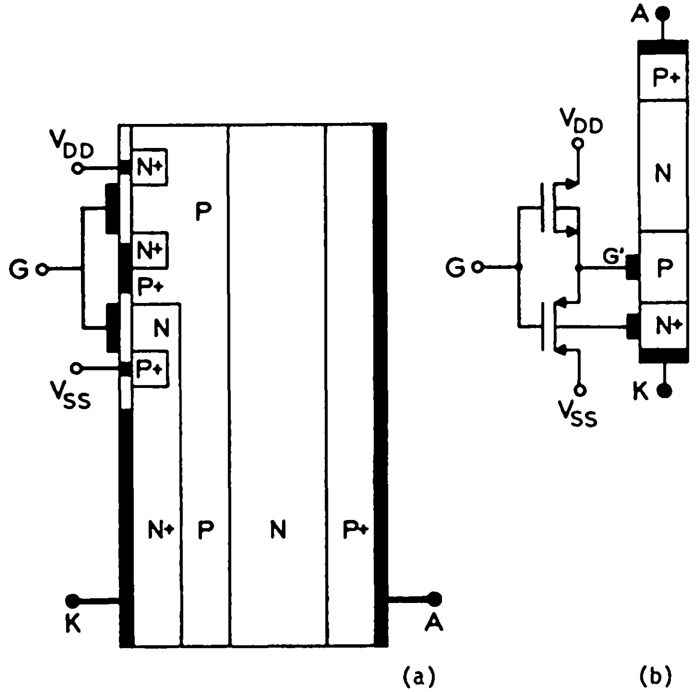
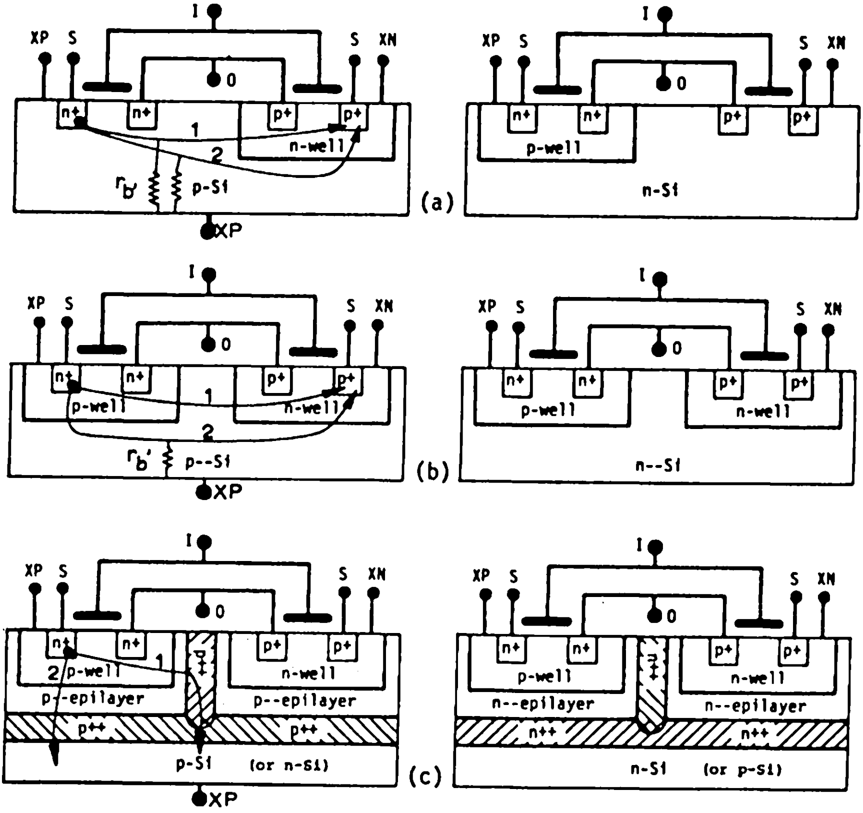


Fig.783.1 A CMOS-SCR. (a) Cross-sectional view. (b) Equivalent circuit.

784 Latch-Up in CMOS

Latch-up in CMOS was qualitatively described in section 672 and Figs.672.23(a)-(c) which are repeated in Figs.784.1(a)-(c). Consider the CMOS' on the left. The base resistance $r_{b'}$ of the p-Si determines the amount of feedback from the p--collector of the p--/n-well/p+ BJT to the p-base of the n-well/p--Si/n+ BJT. If $r_{b'}=0$, then there is no feedback and no latch-up. This simple consideration immediately suggests the solution in figure (c) using epitaxial layer on heavily doped substrate and isolation wall for latch-up prevention. An analysis of the latch-up condition can be made by noting that XP and XN are the two base terminals of the n+/p/n/p+ so that any noise or transient voltages between XP and S_{n+}(positive power supply) or X_N and S_{p+}(negative power supply) could trigger and latch up the CMOS. In addition, $r_{b'}$ in the p-base and a similar resistance in the n-well are emitter shunts if XP and S_{n+} are tied together and X_N and S_{p+} are tied together, so the shunted-emitter SCR analysis given by (782.8) can be used.



Figs.784.1 Latch-up current paths in CMOS. (a) 1-well. (b) 2-well. (c)Epitaxial 2-well.

799 BIBLIOGRAPHY

The following is a chronological list of historical textbooks and reference books on bipolar junction transistors which have been consulted while writing this chapter. The latest articles in conference proceedings were also consulted, such as IEDM (International Electron Device Meeting), ISSCC (International Solid-State Circuit Conference), VLSI Conference, and BCTM (Bipolar Circuits and Technology Meeting). The articles are listed in the sections where they are cited. The following booklist contains only about half of the semiconductor device and circuit textbooks in English that contain bipolar junction transistors. These are selected for inclusion in this list for their historical first, unique, or innovative descriptions and treatments not found in other books. Unfortunately none of the listed books contain all the topics, all the necessary fundamentals, unique and illustrative physics, and examples, which are described in this chapter. Otherwise, there would be no need to write this chapter. The sections of this chapter are written as complete and self-contained as possible. However, the readers and students may still want to consult these books for a comparison of basic physics, analysis details and alternative viewpoints and presentations. (Affiliation of the author is given in the parenthesis.)

[799.1] William Shockley (Bell Telephone Labs. Murray Hill, Shockley Transistor Corp.-Beckman Instruments, Stanford), **Electrons and Holes in Semiconductors**, D. Van Nostrand Company, Inc. New York, 1950. Chapters 2, 4, and 12 are on BJT.

[799.2] Karl R. Spangenberg (Stanford), **Fundamentals of Electron Devices**, McGraw-Hill Book Company, 1957.

[799.3] R.D.Middlebrook (Stanford, Caltech), **An Introduction to Junction Transistor Theory**, John Wiley & Sons, Inc. New York, 1957.

[799.4] Richard F. Shea, editor, (GE Electronics Lab. Syracuse), **Transistor Circuit Engineering**, John Wiley & Sons, 1957.

[799.5] Alvin V. Phillips (Motorola, Phoenix), **Transistor Engineering** and Introduction to Integrated Semiconductor Circuits, McGraw-Hill Book Co. 1962.

[799.6] The seven-volume MIT SEEC series listed on page 829 as [750.1] to [750.7], 1964-66. (Massachusettes Inst. Technology.)

[799.7] Joseph Lindmayer and Charles Wrigley (Sprague Electric Co. North Adams, MA), **Fundamentals of Semiconductor Devices**, D. Van Nostrand Company, Inc. New York, 1965.

[799.8] James F. Gibbons (Stanford), **Semiconductor Electronics**, McGraw-Hill Book Company, 1966.

[799.9] William A. Stover, **TI Series 54/75 Integrated Circuits**, Texas Instruments, Inc. Dallas, 1966.

[799.10] R.L.Pritchard (GE, TI, Stanford), **Electrical Characteristics of Transistors**, McGraw-Hill Book Company, New York, 1967.

[799.11] Charles S. Meyer (Motorola), David K. Lynn (Motorola), and Douglas J. Hamilton (U.Arizona), **Analysis and Design of Integrated Circuits**, McGraw-Hill Book Company, 1968.

[799.12] Paul E. Gray and Campbell L. Searle (M.I.T.), **Electronics Principles, physics, models and circuits**, John Wiley & Sons, 1969.

[799.13] S.M.Sze (Bell Telephone Labs. Murray Hill, Taiwan Chao-Tung U.), **Physics of Semiconductor Devices**, 1st ed. 1969 and 2nd ed. 1981 John Wiley & Sons.

[799.14] Robert L. Morris and John R. Miller (Texas Instruments, Inc.), **Designing with TTL Integrated Circuits**, McGraw-Hill Book Company, 1971.

[799.15] D.J.Hamilton (U.Arizona), F.A.Lindholm (U.Arizona, U.Florida), and A.H.Marshak (U.Arizona, LSU), **Principles and Applications of Semiconductor Device Modeling**, Holt, Rinehart and Winston, Inc. 1971.

[799.16] Jacob Millman and Christos C. Halkias (Columbia U.), **Integrated Electronics: Analog and Digital Circuits and Systems**, McGraw-Hill Book Company, 1972.

[799.17] Victor H. Grinich (Stanford, Berkeley) and Horace G. Jackson(Berkeley), **Introduction to Integrated Circuits**, McGraw-Hill Book Company. 1975.

[799.18] Paul R. Gray and Robert G. Meyer (U. California, Berkeley), **Analysis and Design of Analog Integrated Circuits**, John Wiley & Sons, N.Y., 1st ed. 1977, 2nd ed. 1984.

[799.19] David A. Hodges and Horace G. Jackson (U. California, Berkeley), **Analysis and Design of Integrated Circuits**, McGraw-Hill Book Company, 1st ed. 1983, 2nd ed. 1988.

[799.20] Antonio R. Alvarez, editor (Aspen-Cypress Semiconductor Corp.), **BICMOS Technology and Applications**, Kluwer Academic Publishers, Boston, 1989.

[799.21] David J. Roulston (U. Waterloo), **Bipolar Semiconductor Devices**, McGraw-Hill Book Publishing Company, 1990.

[799.22] George D. Vendelin (Consultant), Anthony M. Pavio (TI), and Ulrich L. Rohde (Compact Software, Inc.), **Microwave Circuit Design Using Linear and Nonlinear Technique**, John Wiley & Sons, 1990.

[799.23] F. E. Gentry, F. W. Gutzwiller, Nick Holonyak, Jr. and E. E. Von Zastrow, **Semiconductor Controlled Rectifiers:** Principles and Applications of p-n-p-n Devices, Prentice-Hall, Inc. Englewood Cliffs, N.J., 1964. See also [799.24] for a collection of earlier papers on p/n/p/n device theory.

[799.24] S. M. Sze, **Semiconductor Devices: Pioneering Papers**, World Scientific Publishing Co., River Edge, New Jersey, 1991.

799 PROBLEMS

P710.1 Determine from the literature the basic reasoning underlying the first analysis of minority carrier injection and diffusion made by Shockley in 1964. How was this different from earlier work by British and German research scientists and book authors?

P720.1 Sketch out the minimum-step process flow and corresponding cross-sectional views of a planar (oxide passivated) BJT with a buried and drift composite collector/base junction which has the LDC (lowly doped collector) structure of n+/p/n-/n+. Use one or all of the five processes: oxidation, diffusion, epitaxial growth, ion implantation, and lithography.

P731.1 By inspection without doing any algebra, write down the d.c. equations of the n/p/n BJT transistor using the two-diode model.

P733.1 Obtain the qualitative criteria on the fundamental material parameters (energy gap, mobility, lifetimes, and dopant impurity concentration) for the Shockley diode current to dominate over the SNS current, under reverse and forward biases, and at low and high temperatures.

P733.2 Obtain the Shockley diode current coefficient, J_1, for the collector-base junction of a n+/p/n/n+ BJT which has a buried n+ collector and whose n-type collector is thin (given by x_c). Let the p/n base/collector junction be located at $x=0$ and the boundary condition at the n/n+ interface be $J_P=0$ or dP/dx. Discuss the result at which the reverse bias is large enough such that the collector/base junction space-charge (depletion) layer becomes equal to the n-layer thickness ($x_{cb}=x_c$), which is known as punch-through.

P735.1 Write down the equations for the reverse active mode by inspection without doing any algebra.

P735.2 Are there any currents flowing in the cutoff mode? What is their magnitude if there are?

P735.3 Obtain the saturation voltage expressions for CB and CE configurations of a n/p/n BJT by inspection of the results listed in the text for the p/n/p BJT. Estimate the magnitude and give the sign following the procedure and assumed parameter values given in the text.

P736.1 Derive the current equation of the underlap diode of a p+/n/p BJT whose base/collector junction depth is X_{BJ} as indicated in Fig.736.1. Assume that the base-metal contact to the n-base surface is ideal (i.e. zero contact resistance or infinite recombination-generation rate). What happens to the current when the collector/base junction bias is so high that its space-charge layer in the base layer side becomes equal to X_{BJ} and explain any limitations on the current?

P736.2 Repeat the above if the base metal contact is offset to a distant region far from the underlap n/p junction and if the surface of the n-base is covered with a SiO_2 which has no recombination-generation centers. What happens to the underlap diode current when X_{BJ} becomes zero? Explain any limitations and the physics.

P736.3 Repeat the above problem for the intermediate case in which the SiO_2/n-base interface has a layer of interface traps whose areal concentration is N_{TT} (traps/cm_2), energy level is at E_T-E_I, and electron-hole trapping coefficients are c_n, e_n, c_p and e_p.

P736.4 The result of the above problem can be simplified using a parameter known as the surface (or interface) recombination velocity for minority carriers in analogy to the lifetime defined in section 37n. Denote this lifetime by S_p for holes in n-base and put your solution of P736.3 in terms of S_p.

P736.5 The lateral base resistance is a fundamental limitation of high-frequency and high-speed performance of BJTs. At low temperatures, deionization of the dopant impurity could increase the base resistance much more than the increase of mobility because of lower phonon scattering at lower temperatures. What is the base resistance at 77K corresponding to the 300K base resistance calculated in (736.6). Take into account each of the following factors individually in a n/p/n transistor and show how much effect each contributes: the deionization of the acceptor dopant impurity in the base layer at low temperatures, the increase of mobility due to less phonon scattering and the decrease of mobility due to more ionized and neutral impurity scattering. Use the result of chapter 3 to calculate the mobility due to ionized and neutral impurity scattering. Since some of these factors improve and others degrade as the temperature is lowered, is there a temperature where r_b, is a minimum and what is it if there is in this example?

P736.6 Repeat the above problem for a p/n/p transistor.

P737.1 The base diffusion delay of a transistor can be significantly reduced by a built-in aiding drift electric field in the base layer via decreasing the majority carrier concentration or the dopant impurity concentration from the emitter boundary to the collector boundary of the base layer. Show from Shockley's current equations that the built-in field is $qE/kT=(d\log_e N/dx)$ in p/n/p and the effective hole diffusivity improvement is proportional to qE/kT. Thus, show that a drop of majority carrier (electron) or dopant (donor) concentration in the base layer by 10 will increase the diffusivity effectively by $\log_e 10=2.303$.

P737.2 Calculate the Gummel number of the base of the two Si n/p/n transistors in Fig.737.1. Make justifiable assumptions about numbers that are not given, some of which can be deduced or roughly extrapolated from the data in section 732.

P737.3 What is the emitter injection efficiency at 77K of the example calculated in (737.15) which has the value of $1\text{-}10^{-3}=0.999$ at 300K? Take into account of the factors described in P736.5 for the lateral base resistance.

P738.1 Suppose that the performance of a BJT transistor is entirely limited by emitter injection efficiency, which limits the d.c. beta, and base spreading resistance, which limits the bandwidth. Show that the gain-bandwidth product is independent of how large the Early effect is in the quasi-neutral base. [Hint: Consider (738.3) and (738.5A).]

P738.2 Present the derivation of the analytical formula for the Early voltage given by (738.7) and (738.7A).

P738.3 How long must the minority carrier lifetime be in order that the SNS current is negligible to give a super-beta Si p/n/p BJT at low currents? Consider the extreme case of beta=10,000 at $V_{EB}=0.1kT/q=2.5\text{mV}$. Use parameter values assumed in the text. If the recombination is entirely due to gold, what must the gold concentration be smaller than, and how many gold atoms are there in the space-charge layer for a area of $100\mu m^2$? Use the recombination data at the gold levels given in chapter 3.

P738.4 The d.c. beta rises with collector current due to the SNS effect and drops due to high injection level in the base. Combine these two effects and obtain the condition at which the beta peaks. Use the simplest formulas derived in the text for these two effects. Find the current density and current of peak beta using the numerical values of the parameters given in this section.

P738.5 Show that the Webster effect cannot account for the rise and then fall of the d.c. beta with collector current such as those shown in Figs.738.2 and 738.3 respectively.

P738.6 Show that under normal conditions, the Kirk effect does not affect the depletion approximation (carrier concentration is zero) at the base edge of the base/collector space-charge layer. Under what condition will the depletion approximation begin to fail because the carrier concentration is becoming comparable to the dopant concentration of the collector. Show numerically that this is at the low or high level injection condition in the quasi-neutral base layer.

P739.1 For the lower voltage and higher speed 2N708 Si n/p/n transistor, the data sheet shows $BV_{CBO}=40\text{V}$, $V_{PEAK}=BV_{CEO}=25\text{V}$, $V_{VALLEY}=19\text{V}$, $I_{VALLEY}=3\text{mA}$, h_{FE} = 11(I_C=0.02mA), 25(0.1mA), 44(1mA), 47(3mA), 50(10mA), 40(100mA), and $I_{CBO}(V_{CB}=20\text{V})=25\text{nA}$. Show that the data are consistent with the negative resistance formulae we have derived in the text. As a parameter extraction exercise, what is the impact multiplication exponent n and the current at peak voltage, $BV_{CBO}=25\text{V}$, to bring these data into a consistent set of results? Answer: n=5.200, $I_{PEAK}=48.44\mu\text{A}$.

P740.1 A electron device has a current voltage relationship of $i=Av^n$ down to $v=0$ where $0<n<1$. What is the condition for small-signal approximation to be valid?

P740.2 What is the small-signal condition for the Fowler-Nordheim tunneling current? See chapter 3 for the forward component of the current.

P740.3 What is the small-signal condition for the band-trap tunneling current? Use the formulae for the forward component of the current in chapter 3.

P741.1 Obtain the analytical expression for the emitter conductance g_e assuming that the emitter quasi-neutral layer is quite thin, i.e., $X_E<<L_E$. Using the approximation of j_E made in obtaining the Gummel number.

P741.2 Suppose the emitter is not so well designed that there is a some thickness modulation of the quasi-neutral emitter layer by the emitter-base voltage due to encroaching of the emitter-base junction space-charge layer into the emitter layer, $X_E = T_E - X_{EB}N_{BB}/(N_{BB}+N_{EE})$ where T_E is the geometrical thickness of the diffused emitter layer. Assuming abrupt junction and constant dopant impurity concentration profiles, i.e., N_{EE} and N_{BB} are spatially constant, what is the formulae of this contribution to g_e computed in P741.1? How much effect does it give when $\gamma_E = 0.95$? (Hint: $X_{EB} = X_{EB}(V_{EB})$ so $\partial X_E/\partial V_{EB} \neq 0$.)

P741.3 Derive (or just copy the expression previously derived for diode) the conductance g_{ebt} due to recombination in the space-charge layer of the emitter-base junction. At what V_{EB} will g_b exceed g_{ebt}? Is this the same injection threshold voltage derived previously in chapter 5 on p/n junction diode and what is the modification and from what origin? (Hint: The base is thin in BJT while in the diode model of chapter 3, a thick-base diode was used.)

P741.4 Obtain the Early voltage, V_{Acb}, from the common-base output characteristics with V_{EB} kept constant, Fig.732.1(a'). Derive the formulae for V_{Acb} using (741.10A). Under what bias condition is it nearly a constant?

P741.5 Result given by (741.9) for the **short-circuit** output conductance of a BJT in the common-base configuration shows that it is very large due to the Early effect conductance, g_a, given by (741.10) which is also verified by the experimental data shown in Fig.732.1(a'). However, the **open-circuit** output conductance in the common-base configuration is very small, mainly from the non-saturated collector junction leakage current due to generation in the space-charge layer, even at very high collector or emitter currents, as indicated by the experimental data shown in Fig.732.1(a). Explain why it is so low and derive the equation to show that it is expected to be so low. [Hint: There are two derivation routes: using the low-frequency CBss-Tee equivalent circuit shown in Fig.731.1(c) or directly from the d.c. CB equations (741.1) and (774.2).]

P741.6 As described before (741.39E), the phase shift and amplitude drop of the small-signal base transport factor of a BJT, $\alpha_b = \mathrm{sech}[(X_B/L_B)\sqrt{(1+j\omega\tau_B)}]$, is well represented by the Middlebrook formula, (741.39B), given by $\alpha_b \approx \alpha_{bo}(1-j0.214\Omega)/(1+j1.04\Omega)$ where $\Omega \equiv \omega/\omega_{\alpha b}$. Show that the amplitude has dropped to $0.707\alpha_{bo}$ at $\Omega = 1$ or $\omega = \omega_{\alpha b}$. Show by Taylor series expansion that at low frequencies when $\omega << \omega_{\alpha b}$, the Middlebrook formula and the 1-pole formula, (741.39A), $\alpha_f = \alpha_{fo}\exp(j\theta_{\alpha b})/(1+j\omega/\omega_\alpha)$ give the same result. What is the phase-shift error in the Middlebrook formula at $\Omega = 1$?

P741.7 A switching delay of $0.254t_\alpha = 0.254(X_B^2/2.43D_B)$ can be obtained from the Middlebrook formula for the base transport factor, $\alpha_b = \alpha_{bo}(1-j0.214\Omega)/(1+j1.04\Omega)$ which was numerically arrived at. This turns out to be very accurate as verified by exact switching transient analysis given in section 75n. Derive this delay by a simple physical argument. [Hint: Compare it with $\alpha_{bo}/(1+j\Omega)$.]

P741.8 The alpha cutoff frequency is changed if a built-in electric field is present in the base. Assume a constant electric field in the base which gives a potential drop of $\eta kT/q$. Show that the minority carrier diffusivity is changed roughly by ηD_B when the field aids the diffusion and D_B/η when the field opposes the diffusion. How accurate are these estimates compared with the accurate formulae for aiding field given by Lindmayer and Wrigly, $1+(\eta/2)^{4/3}$, for with $0<\eta<10$; and by Pritchard, $1+0.080\eta$, for beta cutoff frequency. Compare these also with the exact constant-aiding-and-opposing-field formulae for the bandwidth defined by the CE $\beta(\omega_{bw} \equiv \omega_t) = 1$ which is given by $\omega_{bw} = \omega_B\{(\eta^2/2)/[\eta+\exp(-\eta)-1]\}$ where $\omega_B = 1/t_B = 2D_B/X_B^2$. The normalized built-in potential is defined by $\eta = \log_e[N_{BB}(0_B)/N_{BB}(X_B)]$ which is the potential drop through the base layer in (kT/q) unit and can be positive (aiding field) or negative (opposing field).

P741.9 Instead of the constant built-in electric field in the base assumed in the preceding problem, if the impurity concentration in the base is given by $N_{BB}(x) = N_0\exp(-x/L)$, known as an **exponentially graded base**, show that the equivalent t_B is to be replaced by $(X_B^2/D_B)\{[\eta-1+\exp(-\eta)]/\eta^2\}$. Why is this identical to the constant field solution given in the preceding problem?

P742.1 Calculate the maximum frequency of oscillation of an n+/p/n-/n+ Si BJT fabricated with 0.1μm electron-beam lithography, with the following parameters. $N_{EE}=10^{20}\text{cm}^{-3}$, $X_E=0.1\mu\text{m}$, $N_{BB}=10^{18}\text{cm}^{-3}$, $X_B=0.1\mu\text{m}$, $W_E=0.2\mu\text{m}$, $L_E=1\mu\text{m}$, $\mu_B=\mu_n(\text{at } N_I=10^{18}\text{cm}^{-3})=300\ \text{cm}^2/\text{V-s}$ from Fig.313.5, $X_{Cn-}=0.2\mu\text{m}$, $N_{DDn-}=10^{16}\text{cm}^{-3}$, $N_{DDn+}=10^{19}\text{cm}^{-3}$, $W_C=0.5\mu\text{m}$. Assume that the lateral resistance under the base contact stripe and the base contact resistance are negligible by having two p+ self-aligned stripes diffused or implanted into the n-collector epitaxial layer.

P742.2 Gibbons' formula of the maximum-frequency of oscillation, f_{Max} given by (742.9), was derived using the one-pole approximation to α_f, (741.39A), with the phase-shift neglected. Derive a new f_{Max} using the Middlebrook's one-pole/one-zero alpha approximation. Recalculate the f_{Max} of the very-high-frequency 0.1μm BJT given in P742.1 using the Gibbons-Middlebrook f_{Max} formula just derived.

P742.3 Can a BJT oscillate at a frequency above f_α and f_{t-} and why? Is your answer consistent with the situation in which $r_{b'}C_c$ becomes much smaller than $1/\omega_{\alpha-1}$ and $1/\omega_{t-}$?

P742.4 If f_{Max} is comparable to f_α or f_t, then the Early effect or base thickness modulation, $X_B(t)$, will introduce another circuit element due to reduced α_b at high frequencies. Obtain the analytical expression of this term and estimate the error in the theoretical f_{Max} using Gibbon frequency, (742.9), if Early effect becomes important.

P742.5 The unity-gain frequency or bandwidth of the common-emitter short-circuit-output current amplifier, $|\beta_f(\omega=\omega_t)|=|I_c(\omega_t)/I_b|=1$, the Gibbon frequencies f_{t-} and f_{Max}, and the **maximum frequency of utility** at which the maximum available power gain is unity, f_{u1}, are frequently interchanged in research-engineering articles and textbooks. Derive the formula of f_{u1}. Is it the same as $f_{bw}(\equiv f_t)$ or f_{Max} and why?

P742.6 If the emitter-base small-signal RC time constant $C_{eb'}/g_{eb'}=(C_e+C_{eb't}+C_{b'})\div(g_e+g_{eb't}+g_{b'})=t_{eb'}=1/\omega_{eb'}$, is no longer negligible, then f_{Max} is smaller than (742.9). Then the algebra is exceedingly and excessively complicated which is the reason of neglecting $r_{eb'}$. Instead of carrying out the tedious algebra, an analytical expression of f_{Max} can be derived in terms of $G_{b'c}$ if $G_{b'c}$ is assumed independent of frequency and is the real part of the 'load' admittance, $Y_{b'c}$, which consists of the tuning capacitance C in series with the emitter-base junction admittance $Y_{eb'}=g_{eb'}+j\omega C_{eb'}\equiv g_{eb'}(1+j\omega/\omega_{eb'})$. $G_{b'c}$ is defined and given by $Y_{b'c}(\omega)=[(1/j\omega C)+(1/Y_{eb'})]^{-1}\equiv G_{b'c}(\omega)+j\omega C_{b'c}(\omega)$. Derive a formula of f_{Max} in terms of $G_{b'c}$ and show that f_{Max} is lowered by the loss from $g_{eb'}$. If $\omega_{Max}>>\omega_{eb'}$, then $G_{b'c}\propto g_{eb'}/[1+(C_{eb'}/C)^2]$ and if $\omega_{Max}<<\omega_{eb'}$, then $G_{b'c}/g_{eb'}=(C/C_{eb'})^2(\omega/\omega_{eb'})^2$.

P742.7 By simple physical argument based on the $Y_{b'c}$ circuit, it can be concluded that L can be tuned such that C is minimized. Then, the f_{Max} of (742.9) is the highest possible f_{Max} when C is nearly zero or $C<<C_{eb}$. Show this by simple algebra. Can the condition of the input current source for I_e be used to demonstrate that f_{Max} given by (742.9) is the largest possible value?

P742.8 The BJT can oscillate at a much higher frequency than ω_α because the phase shift due to the transit delay through the base layer can exceed |-90°|. This is known as the **transit time oscillator**, and its frequency of negative resistance or oscillation is increased if there is a large built-in aiding drift electric field. Show this by simple physical reasoning. (Hint: Assumes that the drift field is so large that it dominates over diffusion.)

P742.9 In the transit-time oscillator problem above, the conditions for more than |-90°| phase shift is governed by the CB alpha which is given by a modification of (741.40). If only the base diffusion delay is considered and all other delays are dropped, the modified formula for the base transport factor is $\alpha_b(\omega,\eta)=\exp(\eta_2/)/\{\cosh\zeta+(\eta/2)[(\sinh\zeta)/\zeta]\}$ where $\zeta\equiv[(\eta^2/2)+(X_B{}^2/D_B\tau_B)+j\omega(X_B{}^2/D_B)]^{1/2}$. Show that a negative real part is obtained when $\omega\propto>5D_B/X_B{}^2$ at $\eta=0$ which peaks at $\alpha_b(\omega=10D_B/X_B{}^2,\eta=0)<|-0.13|$, and when $\omega\propto>15D_B/X_B{}^2$ at $\eta=8$ which peaks at $\text{Re}[\alpha_b(\omega,\eta)]<|-0.43|$.

P742.10 Oscillation in the transit time mode has proven experimentally elusive. Why is this so? (Hint: Consider the effects from the neglected losses and capacitances.)

P743.1 Compute the element values of the small-signal hybrid-π equivalent circuit for a transistor whose emitter has an area of $5\text{x}10^{-4}\text{cm}^2$ or $(W_E\text{x}L_E)=100\mu\text{x}500\mu\text{m}$. The doping concentrations and emitter and base thickness are the same as the numerical example given in the text. Are the time constants changed by this scale up in area to give a high power (from high current) transistor and why?

P743.2 There is a conflict of requirement between high-frequency and high-voltage. At what bulk resistivity or collector junction breakdown voltage will the collector series resistance become a limiting factor on the frequency response if the wafer thickness is maintained at $X_C=500\mu\text{m}$ to prevent breakage. (Compute the r_cC_{cj} and make it equal to t_B for a $X_B=1\mu\text{m}$.) Show numerically that a buried n+ collector will give a high-frequency and high-power BJT solution.

P743.3 The transit-time delay of a small-signal passing through the collector/base junction space-charge layer is given by $X_{CB}/2\theta_{sat}$. Show that this is physically correct for the present case of a minority carrier drift through the high-field space-charge layer, and that it is also correct for the case of electrons in a vacuum diode tube. Show especially why this is only one-half of the total transit time X_{CB}/θ_{sat} of an electron passing through a distance X_{CB} at a constant velocity of θ_{sat}.

P751.1 Show that the approximate diffusion equations derived for a constant electric field, (750.3A) and (750.4A), are also valid if E is not spatially constant but the region is quasi-neutral. This enables a derivation of a quantitative criterion on quasi-neutrality.

P751.2 Show that the Kirchoff law used to derive (751.6C) and the charge-control equation for the terminal base current (751.7B) can be rigorously obtained by integrating the hole and electron continuity equations over the three-dimensional volume of the quasi-neutral base. This derivation will show that the terminal base current is indeed the majority carrier (electron) current due to recombination of the minority carriers (holes) injected by the forward biased emitter-base junction with the majority carriers (electrons) in the base layer which are replenished by the electrons flowing into the n-base from the base terminal lead.

P751.3 Suppose that the emitter current in the CB configuration can be defined by a total base charge and the same charge control time constant as the collector current, i.e., $j_{EF}(t)=q_{BF\text{-}TOTAL}(t)/t_{BF}$ while $j_{CF}(t)=q_{BF}(t)/t_{BF}$. Show that $q_{BF\text{-}TOTAL}=q_{BF}+\Delta q_{BF}$ where Δq_{BF} is the small triangular area above the main triangle whose area is q_{BF} in Fig.751.2 and is given by $\Delta q_{BF}=qP_0X_B^3/(4D_B\tau_{BF})$. This gives a geometrical representation.

P751.4 Use the small additional triangular area given in P751.3, show that $dP_0/dT+P_0=\alpha_FP_\infty[1-\exp(-T)]$ which gives an improved charge-control short-time solution of the CB turn-on transient. The agreement with the exact solution of the initial delay (collector current reaches 10% of final current), $T_{0.1}=0.2604$, is further improved if one adds a term to give $dP_0/dT+P_0=\alpha_FP_\infty[1-\exp(-T)+2T\exp(-T)]$. Is the solution good for long times and why is there an overshoot which limits its utility and application at long times?

P752.1 To appreciate the result of the 2x time constant of the collector current decay from capacitance feedthrough when a BJT is driven by a step emitter current, do the following alternative problem. Show directly that if a p/n junction is charged by a reverse current step of 1 unit, which has a very small source conductance, g_s, then the junction voltage increases as $\exp(-2g_st/C_0)$ and the junction current decreases as $1-\exp(-2g_st/C_0)$. Here C_0 is the initial capacitance of the junction and it varies as $\surd\{1+[v_j(t)/V_{bi}]\}$ where V_{bi} is the built-in voltage. What are the current and voltage waveforms if the impurity profile of the p/n junction is graded on one side with an abrupt change to a very large value on the other side.

P752.2 In the preceding transistor problem, suppose that the base resistance is also voltage dependence and is increasing as $1/\surd\{1+[v_j(t)/V_{bi}]\}$ because the base layer is thinning as the collector-base junction is charging up by the reverse current step. Derive and sketch the current and voltage waveforms.

P752.3 If the emitter is floating and a reverse voltage step is applied across the base-collector n/p junction terminals, what are the current and voltage waveforms if $r_{b'} = r_{b'0}\surd\{1+[v_{BC}(t)/V_{CB\text{-}bi}]\}$ and $C_{bc} = C_{bc0}/\surd\{1+[v_{BC}(t)/V_{CB\text{-}bi}]\}$. The charge is given by $q_{BC}(t) = C_{bc0}V_{CB\text{-}bi}\surd\{1+[v_{BC}(t)/V_{CB\text{-}bi}]\}$. [Answer: The response is truly linear although the system or its elements are nonlinear. If the driving force is a constant current source, then the voltage waveform is not linear. Note that this is a way of getting the capacitance-voltage relationship or implementing parameter extracting of the current-voltage-charge versus time relationship from a nonlinear p/n junction.]

P753.1 Derive the exact formula for the short-circuit collector current in the CE turn-on and turn-off switching using the Laplace transform and separation of variable methods for a base current step and a base voltage step. And verify the formulas given in section 753.

P755.1 The effect of built-in electric field on the pulse delay of a BJT is described following (755.7) and estimated using the built-in field factor η (the potential drop through the base layer in units of kT/q). Use the limiting case of large drift field so that drift current dominates in the base. Show that the drift delay is given by $t_{B\text{-}drift} = X_B/\Theta_{drift} = X_B^2/(\eta D_B)$ where Θ_{drift} = drift velocity in the base $= \mu_B E_B = \mu_B(\eta kT/qX_B) = \eta D_B/X_B$. This drift delay is a factor of two smaller than the field-reduced diffusion time used in the text. Thus, the total delay is somewhere between the two solutions. Show that the small-signal formulae of the bandwidth given in P741.8 gives the correct asymptotic limits at $\eta=0$ (diffusion dominates) and $|\eta|$=large (drift dominates): $t_4^{-1}(755.4) = \omega_{bw} = \omega_B\{(\eta^2/2)/[\eta+\exp(-\eta)-1]\}$ where $\omega_B = 1/t_B = 2D_B/X_B^2$.

P756.1 Show that if the number of CE BJT in the ring is even, then the ring circuit will not oscillate since the lowest stable frequency is $\omega_{k=0}=0$. Why is this so physically? Can it be made to oscillate by connecting passive circuit elements between the appropriate nodes?

P756.2 Show that if the CB BJT configuration is connected in a ring, the ring circuit will not oscillate. Why is this so physically? Can it be made to oscillate by connecting passive circuit elements between the appropriate nodes?

P766.1 Describe the digital operation of the EB-shunt reverse-biased drain BiMOS circuit shown in Fig.766.2(a).

P766.2 Solve the d.c. current-voltage characteristics of the EB-shunt forward-biased drain BiMOS circuit shown in Fig.766.1(a) and show with a minimum amount of algebra based on the simplest approximations that the differential transconductance is increased by β_f.

P766.3 Solve the d.c. current-voltage characteristics of the EB-shunt reverse-biased drain BiMOS circuit shown in Fig.766.2(a) and show with a minimum amount of algebra based on the simplest approximations that the differential transconductance is increased by β_f.

P766.4 Figure 766.1(b) shows the emitter up and Fig.766.2(b) shows the emitter down physical layout of the forward and reverse biased EB-shunt BiMOS circuits. Draw the respective emitter down and emitter down physical realizations. [Hint: Emitter up of Fig.766.2(b) requires only minor change of metal interconnect lines.]

P766.5 Figures 766.3(a)-(d) show the emitter-up physical realization of the four fundamental series-BiMOS building block. Give the four corresponding emitter-down physical realizations.

P766.6 The CBiCMOS with CBJT emitter-follower output given in Fig.766.5(d) has the two base nodes disconnected. Can the two base nodes be tied together? Explain whether the inverter still works properly by tracing through one switching cycle.

P766.7 Figure 766.6(a) shows one physical realization of the NBiCMOS with emitter-up BJTs. Give the other three physical realizations of each of the two BiCMOS and two CBiCMOS inverters using emitter-down and combined emitter-up and emitter-down BJTs.

P766.8 In the text, the BJTs in the BiMOS, BiCMOS and CBiCMOS inverters are all vertical transistors. Implement one of each of these circuits using the surface lateral BJTs. (Instructor can assign just one inverter circuit as an exercise to illustrate the principle rather to work out all the possible inverter circuits.)

P766.9 The operation principle of the two full-swing speed-up NBiCMOS inverters in Figs.766.7(a) and (b) were explained in the text by considering the two inverters cascaded and the input=HIGH. Give the explanation like the text when the input=LO is applied to (a) resulting in an input=HI applied to (b).

P766.10 Figures 766.8(a)-(d) gave four NBiCMOS inverters using on-off switched and forward-biased MOST for the EB-shunts to speed up the BJT turn-off transient without attaining rail-to-rail full-swing. Draw the circuit diagrams of the four similar inverters where the EB-shunts are reverse biased.

P766.11 Following the lead of Figs.766.8(a)-(d), give one or more NBiCMOS inverter circuits in which the EB-shunts are not switched off completely so that rail-to-rail full-swing is also attained. Explain with numbers (threshold voltage values) how the EB-shunts and the circuit works to give the full-swing. Is speedup sacrificed or degraded and explain? What parameters need to be controlled to minimize slow-down if speedup is degraded?

P771.1 Draw to scale a tetragonally deformed Si unit cell in an <001> direction with 8 atoms which has $a=b=a_{Ge}$ and $c=K_c a_{Si}$. Draw in as many Si atoms in the unit cell as you can without obscuring the 3-d picture via analytical projection. Determine K_c by assuming the volume is conserved. This illustrates the growth of an Si thin crystalline layer commensurately on a Ge subscript.

P771.2 Repeat P771.1 for a Ge commensurate film on a Si substrate.

P771.3 Repeat P771.2 for a Ge_xSi_{1-x} commensurate film of $x=0.5$ on an Si substrate and use a smaller solid circle for Si+4 core and larger hollow circle for Ge+4 core to simulate approximately the core radius and use $a=xa_{Ge}+(1-x)a_{Si}$ for the equilibrium lattice constant of Ge_xSi_{1-x}.

P771.4 How many layers must one have for a given value of x so that the commensurate film layer of Ge_xSi_{1-x} in the [001] direction is a superlattice? What is the superlattice period? Choose a value of x to give a simple answer, such as $x=0.16666$, 0.25, 0.5, etc.

P773.1 Use a sketch of the aperiodic crystal potential energy of the electrons near the surface of two semiconductors to show that indeed the electron affinity difference (EAD) rule is fundamentally valid. **Be sure to explicitly label the vacuum level.** What is the idealization that makes this rule valid and how is it affected in a real situation? [Hint: Graded or 'reconstructed' heterointerfacial depicted by Fig.774.1(b), (b'), (c) and (c').]

P773.2 Use a sketch of the aperiodic crystal potential energy of the electrons near the interfaces of three semiconductors, A/C/B, to show that the middle (central) semiconductor, C, has no effect on the EAD rule obtained in P773.1 if the middle semiconductor, C, does not strain the interfacial surfaces (or the interatomic spacing of the interfacial atoms) of the two outer semiconductors, A and B. What changes would you expect in E_C and E_V if C does create surface strains in A and B at the interface junctions A/C and C/B?

P773.3 State qualitatively and demonstrate quantitatively the statement in the paragraph after (773.7) that the built-in field comes from $E_G(x)$ in Ge_xSi_{1-x} base layer of the n/p/n HBJT and not due to $E_V(x)$ which has discontinuities. This shows that in spite of the discontinuity of one band edge, the built-in electric field may not come from the band edge that has the discontinuity.

P773.4 Do the same problem as P773.3 for the p-Si/n-Ge_xSi_{1-x}/p-Si HBJT, to show that the built-in electric field comes from $E_V(x)$ and not $E_C(x)$ which is flat.

P773.5 If the $E_G(x)$ in the base due to $Ge_x(x)$ is linear through the base layer of an Si/Ge_xSi_{1-x}/Ge HBJT, what is the diffusivity enhancement factor due to the built-in drift electric field? Assume that $E_G = 1.06$eV for Si and 0.66eV for Ge. Use the previous section on drift electric field in the base to calculate η. Compute this for 300K and 77K.

P773.6 Add to Table 755.1 a fifth BJT which is an Si/Ge_xSi_{1-x}/Si BJT with the same geometrical dimensions as the Si BJT (4) in the table. Assume a $Ge_x(x)$ grading to give a $\Delta E_V = 0.2$eV.

P773.7 Optimize the dopant profiles and magnitudes as well as the drift collector thickness so that the HBJT will exceed 200GHz at the optimum (but quite low) V_{CB} and at 300K.

P773.8 How much improvement on performance can one get if $T = 77$K assuming that the ionized impurity scattering mobility described in chapter 3 and approximated by empirical formula for Si in Table 313.1? Take into account deionization in the base using the figures in section 252. Is the optimum HBJT at 300K of P773.5 also the optimum HBJT at 77K and what needs to be changed to increase the performance at 77K, (i) without making a new transistor (i.e. via changing the operating point), and (ii) by making a new transistor (via changing the material parameters)?

P780.1 Estimate the material properties (i.e. recombination lifetime and base transit time) from the hysteresis loop of the Shockley 4-layer Si diode in Fig.780.1 (60Hz) using the switching transient results of the BJT presented in sections 752 and 753. Explain your numerical result and consider the delays in the V-I curve tracer.

P780.2 How does the alpha of the point-contract n-Ge transistor shown in Fig.780.2(a) vary with collector current and describe the device physics of this variation.

P780.3 Elaborate in more detail the fundamental fallacy of the barrier-lowering space-charge model of $\alpha_F > 1$ in the point-contact transistor.

P780.4 What is the fundamental reason that the emitter-point contact on a n-Ge base must be formed to give good injection efficiency? (Hint: Consider Schottky barrier height for holes.)

P780.5 Why is point-contact Si BJT much harder to make than Ge? (Hint: Schottky barrier heights, impurity diffusivity and temperature, mobility, and recombination lifetime.)

P781.1 Derive a more complete 4-layer diode d.c. V-I equation than (781.1) using the SNS BJT model given by (733.19) and (733.20) or (734.1) and (734.2) by adding the current multiplication factors. Then, relate I_{10} and I_{20} of (781.4) to the basic device material and geometry parameters.

P781.2 Extend P781.1 to high currents by including the high level boundary condition in the base layers.

P781.3 Derive a more complete 4-layer diode d.c. V-I equation than (781.1) by using the extended Ebers-Moll BJT equations given by (734.3) and (734.4). Relate the breakover and holding conditions to the E-M α_F and α_R as well as V_{sat}'s. Assume $M_p = M_n = M_3 = M$ or $\alpha_p = \alpha_n$.

P781.4 Obtain the analytical solutions of the breakover and holding voltages and currents using the SNS model for the current dependence of the alpha used in the text, (781.2), and a power-law voltage dependence of the integrated impact generation coefficient, (781.6A).

P781.5 Why is the linearly graded junction model used to compute the integrated impact generation coefficient in Si presented in Fig.781.3 rather than abrupt step junction? (Hint: Consider the temperature and time required to produce a high voltage p/n/p/n.)

P781.6 Calculate the switch-on delay of a 4-layer diode by a fast voltage ramp assuming that α_{1F} is high due to a thin base and α_{2F} is low due to a thick base.

P782.1 Repeat P781.1 for the standard SCR with a ohmic base contact.

P782.2 Repeat P781.2 for the standard SCR.

P782.3 Repeat P781.3 for the standard SCR.

P782.4 Repeat P781.1 for the shunted-emitter SCR.

P782.5 Repeat P781.2 for the junction-gate SCR.

P782.6 Repeat P781.3 for the remote-gate SCR.

P782.7 Obtain the V-I equation of the bilateral p/n/p/n triode by partitioning the TRIAC of Fig.780.2(g) and apply the 1-dimensional results of the V-I equations for the four gates (ohmic, shunted, junction, and remote) derived in the text to each partition. Verify that the equation gives V-I curves shown in Fig.780.2.

P782.8 Calculate the gate turn-on gain of a emitter-shunted SCR.

P782.9 The transistors of a hypothetical SCR has nearly square shaped α versus I curves. Design a SCR with a turn-off gain of 10 at $I_A/I_H = 10$.

P783.1 Derive the V-I expression of the CMOS-SCR in Fig.783.1. Sketch your results for a range of V_G.

P783.2 A CMOS is used to drive the junction-gate of a SCR, sketch the V-I characteristics. Repeat for shunted-gate and remote-gate.

P784.1 Obtain a simple V-I equation including the base resistance, $r_{b'}$, to estimate the latch-up condition of a CMOS.

P784.2 A linear voltage ramp appears on the n+ source of the CMOS whose substrate is grounded. Derive an expression that can predict the latch-on condition. Include the base-substrate resistance $r_{b'}$.

Appendix A
NOTATION CONVENTION

A judicial choice of notation is crucial to facilitate applications of the seven transport phenomena (drift, diffusion, generation, recombination, trapping or capture, detrapping or emission, and tunneling), and to distinguish the ten participants, (electrons, holes, trapped electrons, trapped holes, electron traps, hole traps, trapping impurity centers, trapping physical defect centers, donor dopant impurities, and acceptor dopant impurities). None of these were considered in the original IEEE Standard on Solid-State Devices developed in 1960. (See Std 216-1960 published by IEEE.) For example, 'diffusion constant' was the recommended term for diffusivity although we know it is not a constant. Thus, extension and revision of the proposed 1960-IEEE Standards on symbols are made which are described in the following paragraphs.

The IEEE Standard of symbols for circuits and devices on a scaler or vector variable is decomposed explicitly into two components to facilitate d.c. analysis, sinusoidal steady-state analysis, linearization for small-signal analysis, and large-signal switching transient analysis. For a scaler variable, the decomposition is

$$v_E(x,y,z,t) = V_E(x,y,z) + v_e(x,y,z,t) \quad \text{scaler} \qquad (A.1)$$

or

$$v_E(r,t) = V_E(r) + v_e(r,t) \quad \text{scaler.} \qquad (A.1A)$$

The time dependence of the scaler variable (the voltage at a node E) and its components given above at a space point $\mathbf{r}=(x,y,z)$ are shown in Figs.A.1(a)-(d) which we shall explain shortly. For a vector variable, to be denoted by a **bold** symbol sometimes underlined for clarity such as the vector field $\underline{\mathbf{a}}_A(x,y,z,t)$ at node A which is located at the point (x,y,z), the decomposition is

$$\underline{\mathbf{a}}_A(x,y,z,t) = \underline{\mathbf{A}}_A(x,y,z) + \underline{\mathbf{a}}_a(x,y,z,t) \quad \text{vector} \qquad (A.2)$$

or

$$\underline{\mathbf{A}}_A(\underline{\mathbf{r}},t) = \underline{\mathbf{A}}_A(\underline{\mathbf{r}}) + \underline{\mathbf{a}}_a(\underline{\mathbf{r}},t) \quad \text{vector.} \qquad (A.2A)$$

Referring to (A.1) and (A.1A), Figs.A.1(a)-(d) show four different waveforms of the voltage at the node E. The upper-case subscript E in the symbol $v_E(r,t)$ or $v_E(x,y,z,t)$ indicates that it is the instantaneous value of the variable 'v'. This instantaneous value may vary with time t as indicated in the four parts of the figure. It may also change spatially which is denoted by space variable (x), (x,y), or (x,y,z) respectively in one-, two-, and three-dimensional cases. And if it is a vector, like that defined by (A.2) and (A.2A), it would have one, two, and three space components given respectively by (1-d) $a_{AX}(x,t)$; (2-d) $a_{AX}(x,y,t)$ and $a_{AY}(x,y,t)$; and (3-d) $a_{AX}(x,y,z,t)$, $a_{AY}(x,y,z,t)$ and $a_{AZ}(x,y,z,t)$.

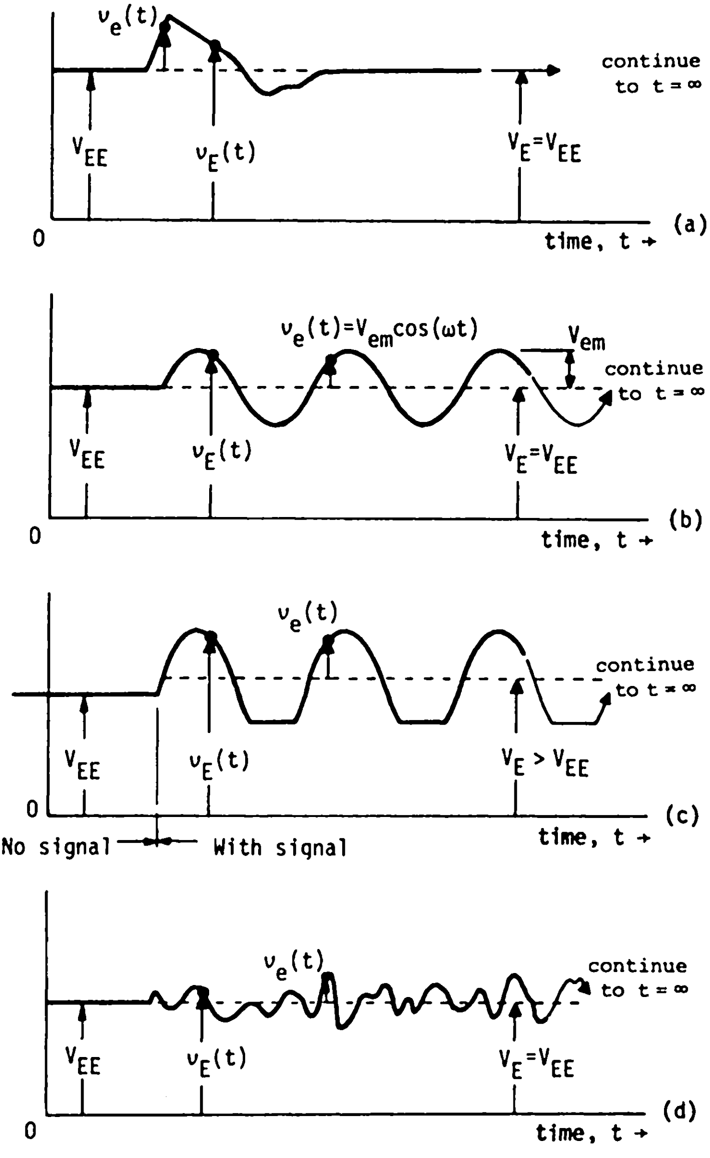

$v_e(t)$
continue to t = ∞
V_{EE}
$v_E(t)$
$V_E=V_{EE}$
0
time, t →
(a)
$v_e(t)=V_{em}\cos(\omega t)$
V_{em}
continue to t = ∞
V_{EE}
$v_E(t)$
$V_E=V_{EE}$
0
time, t →
(b)
$v_e(t)$
continue to t = ∞
V_{EE}
$v_E(t)$
$V_E > V_{EE}$
0
No signal
With signal
time, t →
(c)
$v_e(t)$
continue to t = ∞
V_{EE}
$v_E(t)$
$V_E=V_{EE}$
0
time, t →
(d)

Fig.A.1 Labeled waveforms to illustrate the IEEE notation convention used in this book. (a) A voltage transient superimposed on a d.c. voltage. (b) A pure sinusoidal voltage superimposed on a d.c. voltage. (c) An unsymmetrical periodic voltage (clipped sinusoidal voltage) superimposed on a d.c. voltage. (d) A purely random noise superimposed on a d.c. voltage.

In Figs.A.1(a)-(d), $V_E(r)$ is the **steady-state, stationary, static, quiescent or average value.** It is computed by averaging the instantaneous value, $v_E(r,t)$, over a sufficiently long time interval within which a steady state is attained. The mathematical definition from which the average can be calculated is

$$v_E(r) = \lim_{T\to\infty} \frac{1}{T}\int_0^T v_E(r,t)dt \tag{A.3}$$

where the duration of averaging, T, is sufficiently long so that time-dependent random fluctuations and variations are averaged out or reduced to below a prescribed minimum, for example, 1%, 0.01% or a value depending on the needs of a specific application. It is evident that a true steady-state exist or $V_E(r)$ has a meaningful value only if an average defined by (A.3) can be found. Strictly speaking, a true steady-state is not attainable (Atoms in solids are diffusing even at room temperature; the universe is changing continuously.). All steady-state conditions are quasi (Quasi means approximate but not quite exactly.) steady states, namely a stationary average can be observed in a sufficiently long T, (such as 1 second or 1 hour) so that the microscopic noise is averaged out, but T is much less than the designated total time under consideration (such as the operating life of: a MOS integrated circuit, 10 years; a solar cell, 20 years; or deep spacecraft, >?100 years). And, the same average value is obtained at a later time within the designated total operating time.

The average value may or may not be affected by the presence of a signal as illustrated by Figs.A.1(a)-(d). Fig.A.1(a) shows that the average value V_E is also the d.c. value before the signal is applied, V_{EE}, because the signal is a single pulse which gives a zero average if averaged over a long time. If the average is taken during the presence of the pulse shown in Fig.A.1(a) and over a time interval much less than the pulse duration, then average value will not be a constant and will depend on time or on which portion of the pulse over which the average is made. This is a nonstationary waveform. Fig.A.1(b) shows that the average value V_E is also the d.c. value which is applied simultaneously with the signal, V_{EE}. The reason for $V_E = V_{EE}$ is that the signal is a sinusoidal voltage whose average is zero. However, Fig.A.1(c) shows that the average V_E is greater than the d.c. value V_{EE} because the signal is not symmetrical and has its own d.c. component (in this figure, a clipped sinusoidal voltage which appears in an overdriven stereo amplifier). Fig.A.1(d) shows that the average V_E is again the applied d.c. voltage because the noise voltage is assumed purely random, thus, it has a zero average.

$v_e(r,t)$ is the time-dependent part of $v_E(r,t)$. It is meaningful or not redundant to $v_E(r,t)$ only if a time-averaged or steady-state value, $V_E(r)$, exists like the four cases illustrated in Figs.A.1(a)-(d). Figures (a)-(c) are known as stationary non-random waveforms and (d), a stationary random waveform. Whether non-random (noiseless) or random (noisy), it is not stationary if a time-averaged value cannot be measured repeatedly. Then, the symbol $v_e(r,t)$ is not necessary. And the symbol $v_E(r,t)$ is used and needs not be decomposed into two components.

The amplitude of $v_e(r,t)$ can be large or small. When it is much smaller than a characteristic parameter of the material or device, for example kT/q in solid state devices if $V_e(r,t)$ is a voltage, then it is known as the **small-signal expansion** to electrical engineers, linearization to mathematicians and system theorists, and perturbation to physicists.

If $V_e(r,t)$ has a sinusoidal time-dependent component, such as $v_e(t)=V_{em}\cos(\omega t)$ shown in Fig.A.1(b), it may be further decomposed into a sinusoidal steady-state component and a transient component for a linear or linearized system. Denoting this $v_e(r,t)$ by $a_e(r,t)$, then

$$a_e(r,t) = \tilde{A}_e(r)\exp(j\omega t+j\theta) + a_*(r,t) \qquad \text{(A.4)}$$

or

$$a_e(r,t) = \tilde{A}_e(r)\sin(\omega t+\theta) + a_*(r,t). \qquad \text{(A.4A)}$$

The sinusoidal steady-state component is denoted by $\tilde{A}_e(r)\exp(j\omega t+j\theta)$ or $\tilde{A}(r)\sin(\omega t+\theta)$ and the transient component is denoted by $a_*(r,t)$ which is often denoted by the same symbol as the small-signal or large-signal notation, $a_e(r,t)$ or $a_E(r,t)$, but this will cause confusion. However, the decomposition using the symbol $a_*(r,t)$ given by (A.4) and (A.4A) has not been proposed previously.

The notation of the sinusoidal steady-state component is further simplified by electrical engineers using a complex number notation.

$$\tilde{A}_e(r)\exp(j\omega t+j\theta) = [\tilde{A}_e(r)\exp(j\theta)]\exp(j\omega t) \qquad \text{(A.5)}$$

$$= \sqrt{2}[A_e(r)\exp(j\theta)]\exp(j\omega t). \qquad \text{(A.6)}$$

The complex number $\tilde{A}_e(r)$ is the peak value denoted by V_{em} for the voltage waveform in Fig.A.1(b), and the complex number $[A_e(r)\exp(j\theta)]$ is the RMS (root-mean-square) value of the sinusoidal steady-state component of the signal. Thus,

$$\tilde{A}_a(r) = \sqrt{2}A_a(r). \qquad \text{(A.7)}$$

It is essential that the IEEE Standards of Symbols are used and adhered to so that junior electrical engineers can use their previous circuit knowledge to learn the device physics. In addition, the symbols must be designed to suggest and to help recognize the physics and hence to also help memorize the symbols themselves.

In applying the IEEE Standards to device analysis, we face additional constraints and require extensions in definitions since we have at least six transport phenomena and ten participants in a semiconductor device. First, the variables describing the phenomena and participants are macroscopic, that is, they are already microscopically averaged via ensemble average over many participating particles in a volume element dxdydx at the time t, and via **time average** over many participating particles in a duration dt at the point (x,y,z). The participants may be **fundamental particles** (electrons, protons, atoms, ions) or **quasi particles**, (electrons, holes, traps, and strictly speaking, also atoms, ions and protons) in the volume element dxdydz. Second, the variables are already **microscopically** averaged over **many collision events** (such as scattering and generation-recombination-trapping transitions) in a volume element dxdydz and a time interval dt. In these averages, the volume element dxdydz must be sufficiently large so that it contains many particles (about 10^6 such as that required to store a bit of charge on the capacitor of a DRAM) to limit the fluctuation in the average. And the duration of average, dt, must also be sufficiently large to cover many collision events (about 10^6) again to limit the fluctuation to a prescribed error. The characteristic time of collision or the time of free flight is very short (about 10^{-14}s or less) in electron scattering by lattice vibration or phonons so that dt must be 10^{-12}s or larger. Thus, the **steady-state** concept defined by average over time given by (A.3) is a **macroscopic** average in contrast to the **microscopic** average just discussed.

Traditional notations may be incompatible with the IEEE Standards on Symbols. Certain semiconductor device symbols have been used for four decades after the invention of transistor in 1948, such as symbols without a subscript, for example, the electron (or hole) concentration, n(x,y,z,t). In such cases, we will adopt the traditional usage but make the necessary extensions to the IEEE Standards, such as

$$n(x,y,z,t) = N(x,y,z) + \delta n(x,y,z,t) \qquad (A.8)$$

where the time dependence is denoted by the lower case greek delta, δ. In some cases, a subscript is used to denote the layer in which the electron concentration is referred to, such as

$$n_B(x,t) = N_B(x) + n_b(x,t), \qquad (A.9)$$

then it conforms with the IEEE Standards of Symbols.

INDEX

The indices for the first three chapters are listed as main and sub indices. The indices of the device chapters, starting from chapter 4 on MOSC, are grouped by chapters. For example, indices for MOSC are under the main heading MOSC. Similarly, p/n junction, m/s junction and heterojunction diodes are under the main heading diode; CMOS, MOST inverter, etc. under MOST; BiCMOS, Shockley BJT equations, Ebers-Moll equations, etc. under BJT. Indices begin with abbreviations and acronyms, then nouns, compound nouns and modified nouns. Compound and modified nouns may be listed at two or more locations but not always, such as SRH kinetics which would be listed also under kinetics, SRH. Named terms will always have a main listing and sometimes a sublisting. Author-name is indexed only if appeared in text discussion; see bibliography at end of each chapter.

A

- Ar, A
  - argon,29
- Affinity
  - see electron affinity
- Ampere Law,31
- Angular momentum,35
- Anderson,R.L.,940d
- Auger
  - capture band-trap,286ad
  - recombination interband,278ad
  - see GRTT
  - see also capture
- Avogadro number,171
- Average
  - properties,2
  - ensemble average,237,258,259
  - temporal average,258
  - time average,235,258
- Azimuthal angle,70d

B

- Balance
  - detail,detailed,171
  - power,236
- Balmer series,40
- Band,bands,1ad
  - completely filled,139
  - conduction,98,100,103,109,123,124,141
    - discontinuity,938,940d
  - empty lattice bands,135,136
  - energy,1ad,100
    - see also energy, band
  - valence,98,100,103,109,123,124,141
    - discontinuity,938,940d
- Bardeen,John,269
- Brattain,269
- Basis vector,13
- Bean,John C.,936,951
- Binding energy
  - trapped electrons,165,166
  - trapped holes,165,166
- Binding force,8
- **Bipolar Junction Transistor**,7,701c
  - active mode,743
  - alpha-beta fall off
    - low current,770,968
    - high current,774,968
  - base spreading resistance,754
  - bias effects,765
  - contents,701c
  - collector multiplication,783
  - circuit applications,885
  - cutoff frequencies,744
  - cutoff mode,748
  - d.c. characteristics,712,723a,754a,
  - data, Si npn,718
  - digital inverter,885
    - CE,887
  - diode representation,712
  - diode, underlay non-overlap,754
  - Ebers-Moll Equations,733a
  - Early effects,765
  - fabrication, npn-planar Si,709
  - frequencies, cutoff,744
  - Gibbons' frequency,814
  - Gummel numbers,759,763
  - large signal switching
    - charge-control/diffusion equations, 830,832,833,834,836
    - common base,838
    - common emitter,862
    - comparison CB and CE,
  - lowly doped drain,765
  - material effects,758
  - maximum frequency of oscillation,814
  - negative resistance,783
  - history,704
  - heterostructure,931c
    - HBJT,HBT,931ad
    - commensurate layers,945
    - history,931
    - fabrication of Ge_xSi_{1-x},936
    - operation principles,937
    - energy band,945
    - phonon spectra,945
  - inverters,906
  - ring oscillator,882
  - saturation mode,749

Bipolar junction transistor (continued)
- Shockley BJT equations,730a
- Silicon Controlled Rectifier,962,965,974
- small-signal alpha,808,826,968
- small-signal analysis
  - charge-control,796
  - comparison,793
  - exact,793
- small-signal condition,791
- small-signal equivalent circuits
  - common-base,799
  - common-emitter,817
- SNS BJT equations,731a
- SNS effects,770
- speed-up CE inverter,897
  - turn-on/turn-off overdrive,897
  - Schottky-barrier Bethe diode clamp,897
- two-port network,739a
- two-port representation,739a

Blakemore approximation,201a

Bloch
- Electron,112
- Method,108
- Theorem,112

Bohr
- atom,33
- radius,165,216
- see also hydrogen

Boltzmann
- approximation,178d,201a
- constant,156,171
- distribution,201
- equation,269
- factor,172,174
- relationship,259
- statistics,206
- transport equation,269

Bond,Bonds,1ad
- covalent,9,93
- electron-pair,9,93,96
- energy,8
- metal,9
- model,93
- hetero-polar,93
- hydrogen,9
- homopolar,93
- ionic,9,93
- Joule,9
- Lewis,9
- Pauling
- ruptured,9
- strength,8,93
- van der Wall,9

Bragg reflection,114

Brattain,269

Bravis lattice, fourteen,18d

C

CCD, charge couple device
- see MOSC

CVC, see Chemical vapor deposition

Capture
- auger,286
- cross section,297
- optical,285
- radiative,285
- see also GRTT
- thermal,283

Carrier, carriers
- hot,236,251
- see electron, hole
- see Shockley

Cartesian coordinate,70d

Cellar method,133
- see also Wigner-Seitz

Center
- color,282
- see bound state
- see trap

Charge
- carriers,
  - see electrons, holes, protons
- neutrality condition,188
- state of acceptor,163
- state of donor,163

Chemical vapor deposition,CVD,24,932,933
- LPCVD,low pressure,933

Circuit
- integrated
- semiconductor,2

Classification
- electrical,7ad
- geometrical,6ad
- materials,4ad
  - gas,4
  - liquid,4
  - solid,4
- mechanical,8ad
- purity,6ad

Coherency,239

Collision,159
- see also scattering, trapping
- impact,9

Composition
- uniform,2

Conduction
- band, see band

Conductivity,10,238ad
- extrinsic,253
- impure semiconductor,253
- intrinsic,252,253
  - Si,253
- n-type,160
- p-type,160

- Conservation
- Conservation
  - energy,36,248
  - momentum,270
- Continuity equation,265
- Contact resistance,506
- Continuity equation
  - of current and charge,265ad
- Coordinate,70d
  - azimuthal angle,70d
  - cartesian,70d
- Coordinate (continued)
  - polar angle,70d
  - polar axis,70d
  - spherical,70d
- Correlation
  - length,5
- Correspondence principle,54
- Coulomb
  - force,4,69
  - hypothesis,4,31
  - law,4,31,254
- Crystals
  - atomic density calculation,22
  - growing single crystals,23
    - Czolchralski,28
    - float zone,24
  - lattice translation vector,13
  - Miller indices,16
  - translation vector,13
- Crystal lattices,12ad
- Crystal structures,17ad
  - diamond,19
  - C,Ge,Si,19
  - two-dimensional,14,22
  - three-dimensional,17ad,18d
    - Bravis lattice,18d
    - cubic,15,16
  - zinc blende,20,21
  - wurzite,22
- Crystalline solids,10ad
- Crystallites,6

D

- DRAM,30
  - see MOST
- de Broglie hypothesis,3,33,34
- Debye
  - screening,218
    - Debye-Huckle,218
    - length,219d
  - specific heat,33
- defects,6
  - physical defects,6
- Degeneracy
  - quantum,69ad,200
  - statistical,200
- Degenerate semiconductor,200
- Delocalization concentration,214
- Delocalized
  - see extended
- Density
  - energy,246
  - frequency,246
  - see concentration
- Density of phonon mode,246
- Density of states,176ad
  - effective,177d
- Denuded layer, see SiO$_2$
- Detailed balance,171
- Detrap,165
- Devices,
  - semiconductor,2
  - solid-state,2
- Diffusion,231c,254ad
  - coefficient,255,256
  - constant,255,256
  - current,256,258,260
  - Fisk's law,258
  - length,430
- Diffusivity,255,256
- Diodes
  - four-layer diode,955,961,965
  - see metal/semiconductor diodes
  - see p/n junction diodes
  - see semiconductor heterojunction
  - see tunnel diodes
- Directions
  - Miller indices,16
- Drift,231c,233,236d
  - current,233,238
  - current density,233
  - mobility,239
  - velocity,233
- Dumke,W.P.,932
- Dynamic random access memory
  - see DRAM

E

- E-k diagram
  - see energy band diagram
- E-x diagram
  - see energy band diagram
- Ehrenfest's Theorem,54
- Einstein
  - relationship,255,259
  - specific heat,33
- Electric field
  - built-in,938
  - electron in,45
  - energy level diagram in,45
- Electromagnetic radiation,32,46
  - dual character,32
  - see also light, photon

- Electron,Electrons,1ad
  - affinity
    - difference,940d
    - germanium,944
    - hydrogen,85
    - silicon,98,100
  - concentration,169,174,175,181ad
    - excess,291
    - formulae,192
    - temperature dependence,193a
  - core electrons,89
  - density of state, see Density of state
  - distribution
    - in Si,94
    - temperature dependence,197a,198
  - electric field,in,45
  - energy level diagram,39,42,44
  - hot in Ge,251
  - inner shell electrons,89
  - intrinsic,159d
  - localized electron,203
    - see also trapped electron
  - orbit diagram,39,42,44
  - one-electron model,2
  - one-electron energy,90,92
  - quantum mechanics of,31
  - valence electrons,94
  - wave motion of,31ad
- Electronic materials,2
- Electronic models of semiconductors,93ad
  - band model,97a
    - filling by electrons,102
  - bond model,93a,96
  - one-electron model,2
- Electronics
  - semiconductor,2
  - solid-state,2
  - vacuum,3
- Emission
  - cross section,297
  - impact,286
  - optical,285
  - see also trapping
  - thermal,283
- Energy
  - average kinetic,156ad
  - conservation law,36,248,270
  - change diagram,44,84,85
  - distance diagram,110,111
  - exchange mechanisms,270
  - gap,98,111,114,931d
  - kinetic,36
  - level
    - deep,283
    - diagrams,38
      - neutral Hydrogen atom,39,42
      - negative Hydrogen ion,84
    - shallow,283
- Energy (continued)
  - one-electron,90
  - potential,36
    - excess,119
    - nonlocal,86
  - transition diagrams,44,84,85
  - wavenumber diagram,110,111,123,124,141
- Energy Band
  - filling by electrons,102
  - metals,131
    - Al,131
  - model,97a,108ad
  - nearly free electron model,109
- Energy band diagram
  - metals,131
    - Al,131
  - semiconductors,125
    - AlSb,128
    - CdTe,129
    - CuBr,130
    - GaAs,128
    - GaP,128
    - GaSb,129
    - Ge,128
    - InAs,129
    - InP,129
    - InSb,129
    - Si,100,103,107,126
    - Sn,128
    - ZnS,129
    - ZnSe,129
    - ZnTe,129
  - tight-binding model,117,123,124
- Epitaxy
  - low-pressure,LPE,933
  - molecular beam,MBE,935
- Equilibrium,152,153ad
  - approach to,159
  - atomic,154,155
  - chemical,2,154,155
  - constant,26
  - dynamic,2,152d
  - electrical,2
  - electron,154
  - electronic,154ad
  - electrochemical,154,155
  - mechanical,2,154,157ad
  - hole,154
  - homogeneous,152
  - ionic,154,155
  - reaction constant,209
  - small deviation from,292
  - static,2
  - statistical mechanics,32
  - thermal,2,154,155,161
  - thermodynamic,2
- Esaki,Leo,287
- Excess concentrations,291

F

FET
  see Field-effect-transistor
Face-center cubic,136
Fermi
  -Dirac function,158,170ad
  -Dirac integral of 1/2 order,179
  distribution function,158,172ad,181a
  energy,138,158
  function,158
  level,158,173,ad,181a
    constancy of,261
    intrinsic Fermi level,181
  neutral Fermi level,492
  quasi-Fermi levels,262
  quasi-Fermi potentials,262
  Fermi screening length,219d
  statistics,206
  surfaces,137,138
  -Thomas screening,218
Field-effect-transistors, other,688d
  Current saturation mechanisms,690
  History
    confined channel,691,692
    heterojunction,691,692
    high mobility,691,692
    JGFET,690
      the first,689
      induced surface channel,689
    MESFET,687
    MOSFET,687
      see also MOST
    quantum-well,691,692
  MOS, see MOST
Fisk's law of diffusion,258ad
Fourier
  analysis,111,113
  expansion coefficient,111
  series,111

G

Ge,19,25,128,183,186,245,251,303,944
GeH_4,937
GeSi,SiGe,20
Ge_xSi_{1-x},931,932,936
GRTT,271d
  See generation,recombination,
    trapping and tunneling
Gas constant,171
Gauss Theorem,416
Generation,231c
  see generation,recombination,trapping
    and tunneling
Generation,recombination,trapping and tunneling,3,270ad,271d
  band-trap thermal(SRH),281,284
  band-trap optical,285
  band-trap Auger-impact,270,286a
  collective, plasmon,270,290a
    band-trap,290
    interband,290
  data of rate coefficients,296ad
    Au,gold in Si,299ad,499
    band-trap tunneling Si/SiO_2
      formulae,305
      data,499
    interband optical,300d,301ad
    interband impact,302d,303ad
      GaAs,303
      GaP,303
      Ge,303
      Si,303
      p/n junction,447
    interband tunneling Si/SiO_2
      electron,<100>,<111>,304
      hole,304
  energy,270
  energy exchange mechanisms,270
  generation,284
  interband thermal,273
  interband optical,275
  interband Auger-impact,278,285
  intertrap,286
  lifetime,291ad
  momentum exchange mechanisms,270
  optical,270
  optical capture,285
  phonon,270
  plasmon,270,290a
  radiative capture,285
  recombination,285d
  SRH,281,284
  Shockley-Read-Hall,281
  table,271d
  thermal,phonon,270
  trap occupation factor,297
  trapping,285
  tunneling
    band-trap,288,499
    data of rate,303,305
    elastic,288
    excess current,287,499
    inelastic,289,499
    interband,288
    intertrap,288
    Thompson-Nishida,288

Gibbons,James F.,814,936
Grain boundary,6

H

- HBJT,HBT,931ad
  - DHBJT,932
- Hall,R.N.272,281,294,934
  - see also SRH
- Hamiltonian operator,119
- Harmonics
  - space,13
  - time,13
- Helium,He,29
- Heisenberg's uncertainty principle,33,139
- Heterojunction,931c
  - see HBJT
  - see MOSC
  - see metal/semiconductor
  - see semiconductor/semiconductor
- History
  - quantum mechanics,32
- Holes,1ad,96,139ad
  - concentration,174,175,181ad
    - excess,291
    - formulae,192
    - temperature dependence,193
  - hot,236
  - localized hole,203
    - see also trapped hole
- Homogeneous,152ad
- Houghton,D.C.,935,948
- Hsu,Charles H.,325
- Hydrogen
  - Bohr model,35
  - electron affinity,83
  - Schrödinger solution,69
  - two-electron solution,83
- Hypothesis
  - see DeBroglie, Planck
  - see law
  - see postulate
  - see principle

I

- IEDM, International Electron Devices Meeting,314
- ISSCC, International Solid-state Circuits Conference,314
- Impact,9
- Imperfections,
  - see grain boundary
  - see defects, impurity
  - see dislocation
- Impure
  - classification,6
- Impurity,160ad
  - Au,287,499
  - As,2
  - B,2
- Impurity (continued)
  - Ga,2
  - P,2
  - Sb,2
  - Si,2
  - acceptors,161
  - amphoteric,162
  - bands,214ad
  - carrier screening of,218ad
  - chemical,2,6
  - deep,166,283
  - deionization effects,163ad,203ad, 205a,209a
  - distribution coefficient,26
  - donors,161
  - dopant,11
  - energy band diagram,168
  - foreign,2
  - ionization,163ad
  - isoelectronic,163
  - occupation factor,203,204ad
  - recombination impurity,12
  - segregation coefficient,26,30
  - shallow,166,283
- Imref,262
- Intrinsic
  - absorption,275
    - see optical absorption
  - carrier concentration,181
    - GaAs,184,186
    - GaP,185,186
    - Ge,183,186
    - Si,182
  - optical absorption,275
    - see also GRTT and light
  - radiative recombination,275

J

K

- Katz,L.E.,314
- Kinetic theory,32
- Kinetics
  - SRH,281d,284,294a
  - Hall-Shockley-Read
    - see SRH,GRTT
  - Shockley-Read-Hall
    - see SRH,GRTT
- Kroemer, Herbert,932

L

- LCAO,108d,120ad
- LED
  - see light emitting diode
- Laguerre polynomials,71

Lattice
    direct,13,22
    reciprocal,13
    scattering, see phonon scattering
    vibration,3,11
        see phonon
Law
    Ampere's,31
    Coulomb's,31,234
    Newton's,31,233,234
    Ohm's,251
    see hypothesis
    see postulate
    see principle
    statistical,233
Legendre functions,71
Lifetime,11,291ad
    band-trap,294ad
    high injection level,293
    high-level,293
    interband thermal, optical,291
    killer,12
    low injection level,293,295
    low-level,293,295
    many traps,295
    SRH,294
Light,
    absorption,43,44,275,276
    emission,43,44,275,276
    emitting diode,287
    infrared,41(Paschen)
    momentum,277
    see GRTT
    see optical
    see photon
    vacuum ultraviolet,40(Lyman)
    visible spectrum,40(Balmer)
    wavelength,276
Low injection level,293
Low level,293
Lyman series,40

M

MBE, see epitaxy
MOSC, see Metal-oxide-semiconductor capacitor
MOST, see Metal-oxide-semiconductor transistor

Metal-oxide-semiconductor capacitor,4,311c
    CCD,charge couple device,314
    CV curves
        acceptor deactivation,325
        distortion,323
        graphs,324ad
        hydrogenation,325
        ideal,320ad
        parallel shift,323
        real,323ad

Metal-oxide-semiconductor capacitor (continued),4
    DCV curve,321d
    HFCV curve,313,320,321d,339a
        strong inversion,340
    LFCV curve,320,321d
    MOSC,313d
    MOST gate,313d
    SiO_2/Si interface,316
    acceptor deactivation,324
    acceptor hydrogenation,324
    accumulation,320d
    area,331
    band bending,358
        see surface potential
    band diagram
        see energy band diagram
    bias
        forward,320d,321
        reverse,320d,321
    Boltzmann factor,330
    boron hydrogenation,324
    capacitance
        accumulation,344
        charge-control,329
        charge-storage,329
            differential,329
            interface trap,332
            semiconductor,332
        depletion,338,353
        electron storage,332
        flat-band,344
        high-frequency,339,353
        hole storage,332
        low-frequency,342,
            exact,353
        oxide,320
        see CV curve
        see MOSCV curve
        semiconductor,327
        small-signal,329
        space-charge layer,327
    charges
        bulk,V_{AA},338
        density,areal,332
        interface (trapped),323,327,352
        mobile,329
        oxide (trapped),323,327,352
        space-charge distribution,350,354a,
          361,362,370
    charge control model,325ad
        1-d model,326
    charge control theory,336
        advanced,347
    data
        energy band,insulators,metals,
            semiconductors,356
        oxidation rate,317

Metal-oxide-semiconductor capacitor (continued),4
  Deal,B.E.313
  Debye length
    extrinsic,343
    local,345,348
    metal,359
  depletion
    majority carrier,328
    capacitance,338
  dielectric constant,327
  dielectric permittivity,327
  energy band diagram,350,354a,361,362
    reverse bias,370
    transient,370
  energy band parameters
    (electron affinity,gap, Fermi energy, workfunction)
    insulator,356
    metals,356
    semiconductors,356
  equivalent circuit
    d.c.,337
    exact small-signal,374
    high-frequency,341
    low-frequency,326
    small-signal,337
      depletion,337
  exact low-frequency capacitance,353
  exact small-signal equivalent circuit,374
  fabrication steps,314,315
    die attach (chip bond),319
    header,319
    wire bond (wire attach),319
  flat band,333
  flat band voltage,335
  gate
    area,331
    capacitance,313
    conductor
    metal
    oxide,313
    voltage,349
    voltage shift,323
      negative,323
      positive,323
  Grove,A.S.,313,340
  history,312
  instability,363
    Na^+ ion drift,313,363
    see also transient
  inversion,320d
    strong,328
    time,369
  Katz,L.E.,314
  Moll,John L.,312
  oxidation rate,317a
  oxide charge,352
  oxide film,315,316
  oxide trap charge,352
  permittivity of free space,327
  Philipp,H.R.,314
  potential
    built-in,335,349
    surface,331,351
  small-signal,329
    equivalent circuit,337
  space-charge distribution,350,354a, 361,362,370
  space-charge layer,358
    thickness,359
  surface
    accumulation,320d
    inversion,320d
    potential,331,349
    space-charge layer,358
  Snow,E.H,313
    sodium ion instability,313
  Terman,Lewis M,313
  Terman (Terman-Moll) method,313
  thermal voltage,330
  threshold
    thickness,328
    voltage,340d
  transients
    capacitance,363
      bulk trap,366
      detrapping phase,I,365
      high-frequency,364,365
      interface trap,367
      inversion phase,II,368
      inversion time,369
    current,363,364,369
      see capacitance transient
      phase I,372
      phase II,372
  traps
    bulk,363
      gold,363,369
    oxide,323,352
    interface,323,352
  voltage
    built-in,335
    flat-band,335
    threshold,
  workfunction
    difference,335,351

Metal/semiconductor diode,5,474ad
  Bethe diode,theory,478,483a
  contact resistance,506
  d.c. current-voltage,
    theory,483]
    experiments,489
  energy band diagram,478,486

Metal/semiconductor diode (continued),5
- experimental diodes,489
  - Al/n-Si,490
  - p+/n-Si,490
  - PtSi/n-Si,490
  - W/n-Si,490
  - WSi_2/n-Si,490
- Fermi level
  - neutral,492
  - pinning,492
- history,474
- integrated circuit,497
- Lilienfeld,476
- limiting current,500
- Mott diode,theory,493
- Richardson
  - constant,488
  - equation,487
- Schottky barrier,478
- Schottky-barrier diode
  - see Bethe diode
  - see Mott diode
- Seitz,F.,476
- workfunction
  - difference,482
  - metal,479
  - semiconductor,481

Metal-oxide-semiconductor transistors ,6,521c
- nMOST,532
- pMOST,532
- acronym,525
- aging,547
  - dangling bond,548
  - healing bond,548
  - rupturing bond,548
  - weak bond,548
  - see interface trap, oxide trap
- application, see circuit applications
- body effects,649
  - five examples,651
  - I-V shape,650
  - threshold voltage,649
- bulk charge,545,645ad
  - body effects,649
- characteristics, d.c.
  - analysis,elementary,538
  - conductivity modulation,541
  - current-voltage equation,
    - four-terminal,553
    - parabolic,555
    - three-terminal,554
  - differential equation,551
  - numerical example,558
  - output,529,530,554
  - saturation current,555
    - mechanisms,555
  - transfer,529,530,532,554

Metal-oxide-semiconductor transistors (continued),6
- characteristics, small-signal
  - charge-control capacitances, 560,566
  - distributed model,572,573
  - drain conductance,555,557
  - equivalent circuit,559,564
  - gain-bandwidth product,570
  - frequency,cutoff
    - transconductance,568
  - high-frequency,568
  - numerical example,558,571
  - transconductance,556,557
- characteristics, switching
  - extrinsic delay
    - see charging and discharging capacitors
  - intrinsic delay,577
  - numerical examples,583
  - power-delay product,581
  - see charging-discharging capacitors
- charge
  - bulk,545
  - interface,547
  - oxide,547
- charging-discharging capacitors,583a,585
  - charging,588
  - charge conservation,593
  - charging-discharging cycle time,592
  - charge transferring,592
  - discharging,590
  - fundamental switching equations,586
  - summary,595
- circuit applications
  - archival random access memory,636
  - CMOS
    - cross-sectional views,627
    - d.c.analysis,629
    - history,625
    - latch-up,627,984
    - why?,624
  - current-voltage equation,601
  - DRAM,603
    - basic operation principle,610
    - cell array architecture,608,609
    - cross-sectional views,607
    - definitions,603
    - error,hard,soft,613
    - evolution,635
    - manufacturing history,605,608
    - memory term definition,603
    - refresh,612
  - EPROM,636ad
    - EEPROM,639
    - flash EEPROM,640
    - FEPROM,FRAM,641
    - UV-EPROM,638
  - FRAM,641

Metal-oxide-semiconductor transistors (continued),6
- circuit applications (continued)
  - inverter,614
    - analyses,619,620
    - encyclopedia,614
      - DECMOS
      - DENMOS
      - EECMOS
      - EENMOS
      - RENMOS
    - physical realization,618
    - power dissipation,623
  - PROM,636
  - ROM,636
  - SRAM,632ad
    - circuit,632
    - evolution,635
  - symbol evolution,596
- circuit symbol evolution,596
- conductivity modulation model,beyond,644ad
- classification,530d-533d
  - conductivity type
    - n-channel,532
    - p-channel,532
  - physical origins
    - doped channel,531
    - induced channel,531
- conductivity modulation model,541
- current
  - physics,529
- diffusion current,653ad
  - see subthreshold
- electric field, see high electric field
- fabrication of Si nMOST,534
- failure,547, see aging
- high electric field, voltage
  - generation-recombination-trapping examples,674
    - boron acceptor hydrogenation,677
    - capture-emission,five,675
    - elastic tunneling,ten,676
    - hydrogenation,eight,676
  - mobility,668
    - longitudinal,669
    - transverse,670
- history,524
- instability,547, see aging
- inversion channel,526,
- memory, see DRAM and EPROM
  - DRAM,603ad
  - EEPROM,339
  - EPROM,636d
  - flash EEPROM,640
  - FEPROM,FRAM,641
  - UV-EPROM,236
- narrow gate effects
  - see short channel effects

Metal-oxide-semiconductor transistors (continued),6
- physical structure,526,528
  - body,527
  - channel
    - doped,528
    - electrical length,528
    - electrical width,528
    - induced,528
    - inversion,528
    - length,528
    - thickness,528
    - type,528
    - width,528
  - drain,527
  - gate,527
    - length,527
    - oxide,527
    - width,527
  - oxide
    - field,528
    - gate,527
    - pad,528
  - source,527
- short channel, narrow gate effects,679ad
  - lowly doped drain,679,680
  - three narrow gate effects,683,684
  - three short channel effects,679
    - submicron Si MOST data,679
- small-signal, see characteristics
- subthreshold
  - analysis,657
  - characteristics,653ad
  - current,660
  - definition,656
  - drain voltage dependence,660
  - effective channel thickness,659
  - I_D-V_G,655
  - onset,intrinsic surface,656
  - slope,654
  - subthreshold voltage,656
  - temperature dependence,661
- switching, see characteristics
- threshold voltage,532,549
- traps,662ad
  - interface,549,665
  - oxide,549,662
- voltage, high, see high electric field

Macroscopic,152d
- parameters,268d

Mass action Law,187ad
- generalized,202

Mass, effective,116
Matthiessen's rule,239d
McKay,287,444
McWhorter,A.L.287
Mechanics, statistical,237

Metallic
  semiconductor-metallic transition,215
Meyerson,Bernard S.,932,936ad
Midgap,181ad
Miller indices,16
  direction,16
  plane,16
Mobility, electron and hole
  conductivity,238
  drift,238
  electric field dependence,251
    scatter limited,251
    velocity saturation,251
      Si,electron,hole,251
  temperature dependence,239a
    ionized impurity scattering,240
      Si,electron,hole,250
    lattice scattering,241,246a
      Si,electron,hole,249a
    phonon scattering,241,246a
Model semiconductor,116
Momentum
  conservation,36
    interband Auger-impact,290
  exchange mechanisms,270

N

Neutral Fermi level,492
Neutrality
  charge neutrality condition,188
Newton's Law,31,233,234
Nishida,Toshikazu,288
Noise
  thermal,235
Nonequilibrium,2
  concentration
    electrons,263
    holes,263
  Fermi levels,263
  Fermi potentials,263
  statistical mechanics,31
  see kinetic theory
Notation,992
  see symbol
Nuclear charge
  effective,89

O

Ohm's law,251
One-electron model,2
Optical transitions
  interband,275
  see GRTT
  see light

P

p/n junction diode,5,381c
  barrier,potential
    built-in,407
    diffusion,407
    equilibrium height,404
  breakdown,411,422
    voltages,450
  built-in
    barrier height,407
    potential
    voltage,407
  circuit symbol,420
  current breakdown,441
    impact ionization,444
    interband generation,444
    mathematical,443,448
    mean free path,447
    mechanism,442
  d.c. characteristics,420
  depletion approximation,409,418
  diffusion length,430
  experimental-theory comparison,451
  experimental data,454
  energy band diagram,399,412
    equilibrium,401
    forward bias,427,428,429
      large,503
    reverse bias,424,425
  fabrication,385
    diffusion,385
      junction depth,395
    diffusivity curves,data,389
      Ag,394
      Al,393
      As,393
      Au,394
      B,393
      Cu,394
      Fe,394
      Ga,393
      Mn,394
      O,393
      P,393
      Sb,393
      Zn,394
    Fisk's law,385
  Fermi level
    equilibrium,399
    intrinsic,403
    quasi-, see quasi
  Gaussian theorem,416
  history,382
  limiting current,502
  Poisson-Boltzmann equation,409
  potential distribution,equilibrium,408
  quasi-neutrality,412

- **p/n junction diode (continued),5**
  - reverse breakdown, see current breakdown
  - Sah-Noyce-Shockley diode equation (SNS),436
  - see also metal/semiconductor
  - Shockley diode equation,430
    - numerical example,436
    - physics,432
  - silicon p/n diode,458
  - small-signal
    - charge-control circuit,456
    - numerical examples,458
  - SNS, see Sah-Noyce-Shockley
  - space-charge layer,411
  - switching transients,461ad
    - charge-control,462
    - turn-on transient,464
    - turn-off transient,467
  - transients,reverse,470ad
    - capacitance,470
    - current,470
    - Tasch,A.F.,470
    - Yau,L.D.,470
  - transition layer,411
    - see space-charge layer
  - tunneling,423,499

- Particles, see also electrons, holes, phonons
  - quasi-particles,4
- Particle distributions,4
- Paschen series,40
- Pauli's exclusion principle,33
- People,Roosevelt,935,949,951
- Periodic structures,12ad
- Perturbation method,119
- Phonon
  - branch,244
  - density of mode,246
  - dispersion spectra, see phonon spectra
  - energy,245d,246d,248
  - frequency,246
  - frequency gap,246
  - mode,244d
  - scattering,240,241,246,ad
  - spectra,241
    - 1-d monoatomic,242
    - 1-d diatomic,242
    - aSiO_2,245
    - GaAs,245
    - Ge,245
    - Si,245
    - ZnSe,245
    - acoustical,244d
    - optical,244d
  - temperature,246d
  - wave
    - longitudinal,242,243ad
    - transverse,242,243ad
- Photoelectric effect,33
- Photon, see GRTT,light
- Planck's condition,hypothesis,3,33,34
- Plane
  - Miller indices,16
- Plasmon recombination,290
  - see GRTT, collective
- Poisson equation,269
- Poisson-Boltzmann equation,409
- Polar angle,70d
- Polar axis,70d
- Poly, see Poly Si
- Poly Si,6
- Polycrystalline,6
- Polycrystalline Si
  - See poly Si
- Postulate
  - de Broglie,3
  - Planck,3
- Potential
  - Coulomb,112
    - see Coulomb
  - electric,174d
  - excess,119,120
  - macroscopic,174
  - periodic,11,109,118
  - see energy, potential
- Power balance,236
- Proton,
  - energy level diagram,45
  - see hydrogen atom, Bohr

Q

- Quantum
  - condition, see Planck
  - multi quantum well,MQW,933,935,937
  - efficiency,275
  - numbers,71
    - angular,71
    - azimuthal,71
    - magnetic,71
    - orbital,71
    - principle,71
  - see also light
  - see also phonon

- Quartz
  - boat,29
  - diffusion tube,29
  - fused,29
  - tube,29
- Quasi
  - Fermi levels,262
    - see also nonequilibrium Fermi level
  - Fermi potentials,262
  - neutrality approximation,411
  - particles,4

R

Radiative transitions, see GRTT,light
Randomness,239d
Recombination,231c, see GRTT
Resistivity
  classification of solid,8
Richardson
  constant,488
  equation,487

S

$SiCl_4$,silicon tetrachloride,24
SiH_4,silane,24,937
  see also Gelane,GeH_4,937
$SiHCl_3$,trichlorosilane,24
SiO,silicon monoxide,29
SiO_2,silicon dioxide, see also MOSC,MOST
  cluster,29
  denuded layer,30
  film,
  Katz,L.E.,314
  optical properties,
    Philipp,H.R.,314
Scattering,3
  impurity
    ionized,240
  lattice,240,241
  mean free path,447
  mean free time,256
  phonon,240,241
  probability,256
Schottky barrier,478
Semiconductors
  $Al_xGa_{1-x}As$,932
  AlN,20
  CdS,20
  CdTe,20
  GaAs,2,20
  GaN,20
  GaP,20
  Ge_xSi_{1-x},931ad,932,936a
  InAs,20
  InGaAs,932
  InP,20,932
  InSb,20
  SiC,20
  SiGe,20,see Ge_xSi_{1-x}
  ZnO,20
  ZnS,20
  ZnTe,20
  binary compound,162
  compound semiconductor,162
  cubic,20
  degenerate,200
  diamond,19
Semiconductors (continued)
  electronic models,93a
    bond model,93
    band model,97
  extrinsic,187,190ad
  -metal transition,214
  model semiconductor,116
  hexagonal system,20
  hexagonal close-packed,20
  impure,10ad,160ad,187
  intrinsic,159,194ad
  pure,10ad,159ad
  wurzite,20,22
  zinc blende,20,21
Semiconductor/semiconductor
 heterojunction diode,510
  electrical characteristics,513
  energy band diagram,510,511,513
  trapless interface,510,513
  trappy interface,512,513
Schrödinger equation,31a,33,46a
  derivation,47a
  see wave equation of electron
Screening
  Debye-Huckle,218
  Fermi-Thomas,218
Seitz,F.,476
Shockley,Shockley's,Shockley-
  book,269
  density,298
  diode current,941
  diode equation,430
  electron affinity difference,941
  equations,268ad,269
  hot electron theory,251
  imref,262
  -Pearson FET experiment,541
  quasi-Fermi levels,262
  Shockley-Read-Hall kinetics,281,284d,285
Solubility,Solid solubility,29
Sommerfeld approximation,201
Space harmonics,13
Spectra,pair,286
  GaP,287
  Si:Li,In,287
  ZnS,287
Spherical coordinate,70d
Spherical harmonics,70
Standing wave,115
Statistical mechanics,237
Sun,Y.C.(Jack),325
Superlattice,6
Symbol
  choice,272,appendix-A,992
  GRTT,272
  IEEE Standard,appendix-A,992
Sze,S.M.,314

T

Tasch,A.F.,470
Temperature,
  electron,156
  intrinsic,194,195d,196a
  lattice,156
  phonon,246
  spatial constant,2
  temporally constant,2
Thermal noise,235
Thermal velocity,235
Thermodynamic equilibrium, see Equilibrium
Thompson,Scott E.,288
Tight-binding model,117
Time
  free time,234,257
  mean free time,235
  relaxation time,237,258,259
Time harmonics,13
Transistors, see BJT,MOST,HBJT,FET
Transition
  energy band diagram,274,276,279,283,286,288,289,290,299,304,305
  processes,171
Trap,trapped,trapping,163,231d,285d
  electron,164,267
  hole,164
  isoelectronic,162
  occupation factor,297
  see GRTT,capture,detrap,emission
Tunnel,tunneling,118,231c, see also GRTT
Tunnel diode,499ad
Tzou,J.T.,325

U

UHV,ultra-high vacuum,933
Unit cell,13
  body-centered cubic,18,132
  face-centered cubic,15,18,19,136
  nonprimitive,14,19,132
  primitive,13,19,132
  volume,23
  Wigner-Seitz,15,19,132

V

Vacuum
  electronics,3
  level,98,109,110
  vacuum/Si interface,98
Vacuum electronics,3
Valence
  band,98,100,103,109
  bandwidth,98,100,103,109
  bond,94
  electron,94
Vectors,
  basis,13d
  primitive basis vector,13d
  primitive translation vectors,13d
  translation vectors,13d
Velocity
  electron,139,141
  group,139
  scatter limited,236
  saturated,236
  thermal,235

W

Wave
  electron, see wave equation,wavefunction
  lattice vibration wave, see phonon
  length,wavelength,276
  longitudinal, see phonon
  transverse, see phonon
Wave equation of electrons,31ad,46a
  experimental bases,46
  general properties,53,56
  solutions
    bound states,66a
    general,56a
    hydrogen atom,69a
    many-electron atoms,87a
    reflection at potential step,57a
    resonance scattering,60a
    tunneling
      square barrier, 62a
      triangular barrier,63a
    two-electron hydrogen,83a
      see hydrogen
Wave function, wavefunction
  classical-quantum connection,53
  interpretation,53
  properties,53
Wave motion of electrons,31ad
  dual character,32
Wavefunction, see wave function
Wavelength,276
Wigner-Seitz cell,15,132
Wurzite structures,20,22

X

x-ray,9

Y

Yau,L.D.,269

Z

Zinc blende structure,20,21

Chih-Tang Sah is the Pittman Eminent Scholar and a Graduate Research Professor at the University of Florida since 1988. He was a Professor of Physics and Professor of Electrical and Computer Engineering, emeritus, at the University of Illinois at Urbana-Champaign where he taught for twenty-six years and guided 40 students to the Ph.D. degree in electrical engineering and in physics. He has published about 250 journal articles and given 100 invited lectures in China, Europe, Japan, Taiwan and the United States on transistor physics, technology, and evolution. He received two B.S. degrees in 1953, in Electrical Engineering and Engineering Physics, from the University of Illinois, and the M.S. and Ph.D. degrees from Stanford in 1956. His doctoral thesis research was on traveling-wave tubes under the tutelage of Karl R. Spangenberg. His industrial career in solid-state electronics began with William Shockley in 1956, and continued at the Fairchild Semiconductor Corporation in Palo Alto from 1959 to 1964 until he became a professor of physics and electrical engineering at the University of Illinois in 1963. Under the management of Gordon E. Moore, Victor H. Grinich, and Robert N. Noyce at Fairchild, Sah directed a 65-member team on the development of the first generation silicon bipolar and MOS integrated circuit technology including oxide masking for impurity diffusion, stable Si MOS transistor, the CMOS circuit, origin of the low-frequency noise, the MOS transistor model used in the first circuit simulator, thin film integrated resistance, and Si epitaxy process for bipolar integrated circuit production. For contributions in transistor physics and technology, he received the Browder J. Thompson best paper prize for an author under thirty, the J. J. Ebers Award in Electron Devices, and the Jack Morton Award, all from the IEEE, the Franklin Institute Certificate of Merit, the first Achievement Award in High Technology from the Asian American Manufacturer Association, and the Doctor Honoris Causa degree from the University of Leuven, Belgium. He was listed in a survey by the Institute of Scientific Information as one of the world's 1000 most cited scientists during 1965–1978. He is a fellow of the American Physical Society and the IEEE, and a member of the U.S. National Academy of Engineering.

Copyrighted Material

Fundamentals of Solid-State Electronics was successfully used as the textbook for the introductory junior electrical engineering core course in solid-state devices taught to about 300 students in six semesters at the University of Florida, which was also attended by undergraduate and graduate students in physics, science, and other engineering departments. It has three constituents, (1) electronic materials physics in three chapters and four device chapters (MOSC, p/n-m/s-ohmic diodes, MOST-FETs, BJT-HBJTs, and SCRs) each contains: (2) history, fabrication, characteristic physics, and circuit models, and (3) basic-building-block circuits. The extended coverage in the second part of each chapter can be selected for a second course and as a reference for practicing engineers and managers on the fundamentals of materials and device physics, device models, and basic-building-block circuits (B^3C) for complex integrated circuits. Examples are: advanced device physics (deionization and heavy doping effects, subthreshold current, high field mobilities, reverse capacitance and current transients in MOSC and p/n and m/s junctions, ohmic contacts, ...), latest state-of-the-art (1990–1991) device concepts (heterostructure MOSFET and BJT, ...), reliability mechanisms (channel hot electron injection, Fowler-Nordheim tunneling, interband hot hole generation and injection; p-Si gate 1.2eV less reliable than n-Si gate, ...), and B^3C (BiCMOS, CBiCMOS, DRAM, SRAM, UV-EPROM, flash-EEPROM, and FRAM). The book develops the fundamental concepts needed in device physics (electrons and holes, bond and band models, equilibrium and nonequilibrium, statistical distributions, drift and diffusion, generation-recombination-trapping and tunneling) from freshman chemistry and sophomore physics (Newton, Coulomb, Planck, and de Broglie Laws). It gives work-out examples which are physically significant and numerically illustrative of the state-of-the-art, such as the submicron Si MOSFET and Si BJTs. It contains nearly 100 selected and critiqued intermediate and advanced book references and about 500 problems for extending the study beyond this book.

1388 hc

ISBN 981-02-0637-2

Copyrighted Material